JMIR Medical Informatics

Impact Factor (2022): 3.2
Volume 10 (2022), Issue 10 ISSN 2291-9694 Editor in Chief: Christian Lovis, MD, MPH, FACMI

Contents

Viewpoint

Original Papers

Viewpoint

Lifting Hospital Electronic Health Record Data Treasures: Challenges and Opportunities

Alexander Maletzky[1], PhD; Carl Böck[2], PhD; Thomas Tschoellitsch[3], MD; Theresa Roland[4], PhD; Helga Ludwig[4], MSc; Stefan Thumfart[1], PhD; Michael Giretzlehner[1], PhD; Sepp Hochreiter[4], PhD; Jens Meier[3], MD

[1]Research Department Medical Informatics, RISC Software GmbH, Hagenberg, Austria

[2]JKU LIT SAL eSPML Lab, Institute of Signal Processing, Johannes Kepler University, Linz, Austria

[3]Department of Anesthesiology and Critical Care Medicine, Kepler University Hospital GmbH, Johannes Kepler University, Linz, Austria

[4]ELLIS Unit Linz, LIT AI Lab, Institute for Machine Learning, Johannes Kepler University, Linz, Austria

Corresponding Author:
Alexander Maletzky, PhD
Research Department Medical Informatics
RISC Software GmbH
Softwarepark 32a
Hagenberg, 4232
Austria
Phone: 43 7236 93028406
Email: alexander.maletzky@risc-software.at

Abstract

Electronic health records (EHRs) have been successfully used in data science and machine learning projects. However, most of these data are collected for clinical use rather than for retrospective analysis. This means that researchers typically face many different issues when attempting to access and prepare the data for secondary use. We aimed to investigate how raw EHRs can be accessed and prepared in retrospective data science projects in a disciplined, effective, and efficient way. We report our experience and findings from a large-scale data science project analyzing routinely acquired retrospective data from the Kepler University Hospital in Linz, Austria. The project involved data collection from more than 150,000 patients over a period of 10 years. It included diverse data modalities, such as static demographic data, irregularly acquired laboratory test results, regularly sampled vital signs, and high-frequency physiological waveform signals. Raw medical data can be corrupted in many unexpected ways that demand thorough manual inspection and highly individualized data cleaning solutions. We present a general data preparation workflow, which was shaped in the course of our project and consists of the following 7 steps: obtain a rough overview of the available EHR data, define clinically meaningful labels for supervised learning, extract relevant data from the hospital's data warehouses, match data extracted from different sources, deidentify them, detect errors and inconsistencies therein through a careful exploratory analysis, and implement a suitable data processing pipeline in actual code. Only few of the data preparation issues encountered in our project were addressed by generic medical data preprocessing tools that have been proposed recently. Instead, highly individualized solutions for the specific data used in one's own research seem inevitable. We believe that the proposed workflow can serve as a guidance for practitioners, helping them to identify and address potential problems early and avoid some common pitfalls.

(JMIR Med Inform 2022;10(10):e38557) doi:10.2196/38557

KEYWORDS

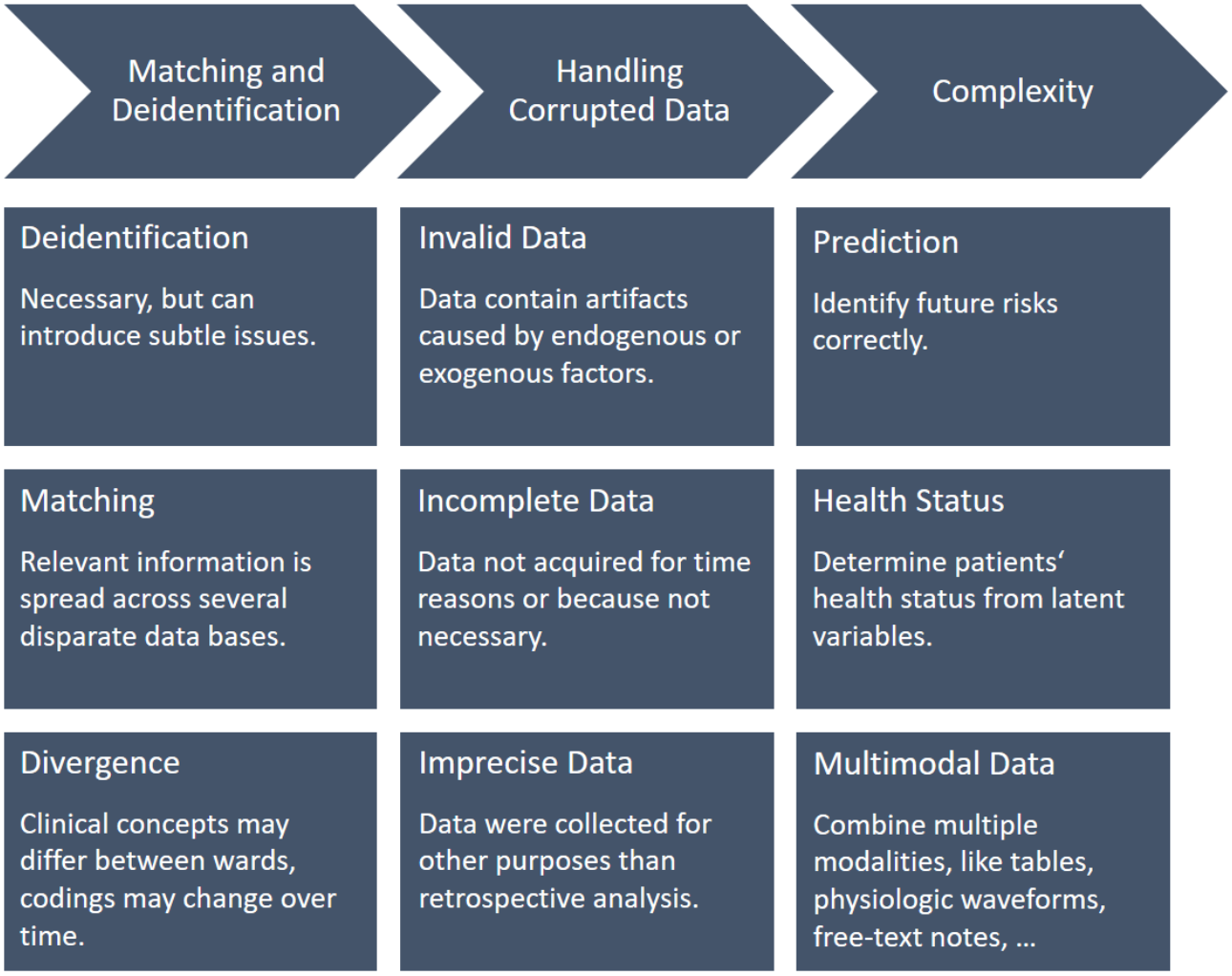
electronic health record; medical data preparation; machine learning; retrospective data analysis

Introduction

Electronic health records (EHRs) contain a vast amount of information about an individual's health status, including demographics, diagnoses, medication prescriptions, laboratory test results, high-frequency physiologic waveform signals, and others. Many prior studies have demonstrated how data science and machine learning (ML) can be applied to large databases of EHRs to successfully train models to predict many different patient-related outcomes, including mortality risk [1-4], length of hospital or intensive care unit (ICU) stays [1-3], cardiovascular decompensation [3,5,6], postoperative complications [7], and, recently, COVID-19 diagnosis and pathogenesis [8-12]. Although data preparation requires

considerable time and effort [13,14], it is seldom represented in research outputs. One possible explanation could be that it is considered a "standard" task that always proceeds more or less the same and that can be automated to a large extent, thanks to readily applicable general purpose software tools [15-17]. In this paper, we illustrate through specific examples from a large-scale research project that this is not the case. Conducting a secondary (ML-based) analysis of raw EHRs from a hospital's data warehouse is challenging in many respects due to several reasons. Above all, data were originally collected without any specific use case besides clinical application, and relevant information is usually distributed over multiple disparate databases that often lack comprehensible documentation. If clinical concepts (variables, categorical values, units of measurement, etc) are represented differently across distinct sources or if the coding of clinical concepts changes over time, data harmonization can become a real issue. Moreover, incomplete or invalid data, although a well-known problem in principle, can occur in many (unexpected) forms and might only be noticed after careful manual inspection. Figure 1 summarizes the main challenges with EHR data that we encountered in our work and are ubiquitous in retrospective medical data analysis [18].

Figure 1. Primary challenges with retrospective medical data analysis (adapted from Johnson et al [18], which is published under Creative Commons Attribution 4.0 International License CC-BY 4.0 [19]).



Unlike many other papers about data preparation in a medical context, this work does not propose a novel generic data processing tool. Instead, we report the challenges that we faced and the lessons we have learned in a recent large-scale data science project. We present specific examples of messy and corrupted raw data to create awareness that (medical) data preparation is a nontrivial, labor-intensive endeavor, despite an ever-growing set of generic tools. Finally, we present a general data preparation workflow for similar research projects to help practitioners avoid the most common pitfalls.

The literature on medical data preparation for large-scale secondary (ML-based) analysis is scarce. Most studies have focused on model development and the final predictive performance of the developed models and only mention a few fundamental aspects of the data preparation pipeline. This is particularly true for the work of Rajkomar et al [1], but for a good reason: deep neural networks are used to learn "good" representations of the data in an end-to-end fashion, relying on the networks' ability to automatically handle messy data properly. The pipeline is based on Fast Healthcare Interoperability Resources [20], meaning that all data available in this format can be readily processed without further ado—no feature selection, harmonization, or cleaning is necessary. Although appealing at first glance, the proposed approach has

some limitations, as noted by the authors. Most importantly, deep neural networks often require massive amounts of data and computing resources to learn good representations. Second, the lack of data harmonization potentially impairs transferability across research locations; for example, for validation. Moreover, to train models in a supervised manner, one must provide labels, and depending on the use case, these labels may be difficult to extract, reintroducing the need for data preparation. It is also unclear whether the models developed in the aforementioned study [1] would have performed even better, had the data undergone more thorough manual inspection and curation.

Other studies have proposed generic data processing pipelines that can be applied off-the-shelf to well-known ICU benchmark databases such as Medical Information Mart for Intensive Care (MIMIC) [21-23] and the telehealth ICU collaborative research database (eICU-CRD) [24]. The most prominent examples being MIMIC-Extract [15], FIDDLE [16], cleaning and organization pipeline for EHR computational and analytic tasks [17], and Clairvoyance [25]. The authors of FIDDLE and Clairvoyance claim that their systems are sufficiently general to accommodate not only data extracted from MIMIC-III and eICU-CRD but also any EHR data available in a particular form. This may be true to a large extent, but we experienced that cleaning messy, raw data and bringing them into the required standardized form is at least as labor intensive (in terms of implementation effort) as the subsequent "generic" preprocessing steps that FIDDLE and Clairvoyance cover. Sculley et al [14] termed this phenomenon *glue code antipattern*. In general, MIMIC and eICU-CRD may be excellent benchmark databases, but we found that "real-world" data exported directly from a hospital's IT infrastructure pose many challenges that are not present in these databases.

Shi et al [26] presented a medical data cleaning pipeline that explicitly addresses some of the issues that we also encountered in our research. They considered laboratory tests and similar measurements and proposed manually curated validation rules for numerical variables and an automatic strategy for harmonizing (misspelled) units of measurement through fuzzy search and variable-dependent conversion rules. The focus of Shi et al [26] is on improving the quality of data [27-29], whereas Wang et al [15], Tang et al [16], and Mandyam et al [17] are mainly concerned with transforming data into a form suitable for ML. A more detailed evaluation of FIDDLE, MIMIC-Extract, and cleaning and organization pipeline for EHR computational and analytical tasks and the approach to our data by Shi et al [26] can be found in Multimedia Appendix 1 [15-17,26].

The extensive survey article by Johnson et al [18] summarizes the main issues of medical data analysis similar to that in this work. The authors also established a high-level categorization of these issues into *compartmentalization*, *corruption*, and *complexity* (Figure 1) and argued that data acquisition and preparation in the critical care context are particularly difficult because data are collected for a different purpose.

Sendak et al [30] arrived at similar conclusions, noting in particular that solutions developed for one site did not scale well across multiple sites because of redundant data validation and normalization. The authors provided estimates for the expected cost of deploying a model to screen patients with chronic kidney disease in other hospitals. We refrain from extrapolating such estimates from our findings but agree that the costs for preprocessing data from other sites into a form suitable for existing prediction models will likely be significant.

Methods

Data Preparation

Raw EHRs stored in hospitals' data warehouses cannot readily be used for developing clinical prediction models but must first be extracted, analyzed, and subjected to a series of preprocessing steps. These steps may differ between data modalities and sources but usually include some sort of *validation* (ensuring data accurately reflect reality), *harmonization* (establishing uniform representation of equivalent concepts), and *transformation* (bringing data into a form suitable for model development, eg, extracting useful information). Furthermore, it must be ensured that a sufficient number of data points are available in the first place and that clinically meaningful target labels can be extracted from them in the case of supervised learning. We demonstrate how this can be accomplished in a disciplined, effective, and efficient manner by referring to a specific data science project.

Underlying Data Science Project

All results presented in this paper originate from a large-scale data science project for developing data-driven clinical prediction models. Specifically, the following 5 use cases were considered: (1) optimizing patient throughput in the ICU, (2) increasing the accuracy of treatment priorities in emergency medicine, (3) improving the selection of blood products, (4) predicting patient deterioration in the ICU to enable preventive interventions, and (5) predicting COVID-19 infections using routinely acquired laboratory tests [11]. All use cases were based on retrospective, routinely collected data from the Kepler University Hospital, a large university hospital in Linz, Austria. A wide variety of data modalities were used, including patient demographics, laboratory tests, diagnoses, vital signs, and even high-frequency physiological waveform signals. Information represented by natural-language text was mostly ignored (except for short free-text diagnoses), and imaging modalities were excluded altogether.

The amount of data varies among the 5 use cases; for instance, use cases 1 and 4 are naturally confined to patients admitted to the ICU, whereas for use case 2, only patients who visited the emergency department (ED) could be taken into account. The period covered by the data also depends on the use case. Table 1 lists the particular period and the total number of patients for each of the 5 use cases. Altogether, the order of magnitude of the number of data items processed was 10^9 (excluding high-frequency waveform data) of which vital signs and laboratory tests constituted the vast majority.

The specific results of the 5 use cases were not the main focus of this paper. Instead, the use cases serve merely as illustrative examples throughout the remainder of this paper.

Table 1. Use cases considered in the research project[a].

Use case	Short description	Period	Patients, n
1	Optimizing ICU[b] patient throughput	2010-2020	14,236
2	Increasing the accuracy of treatment priorities in ED[c]	2015-2020	77,972
3	Improving the selection of blood products	2016-2020	5855
4	Predicting patient deterioration in the ICU	2018-2020	3069
5	Predicting COVID-19 infections [11]	2019-2020	79,884

[a]Note that patient cohorts partly overlap.

[b]ICU: intensive care unit.

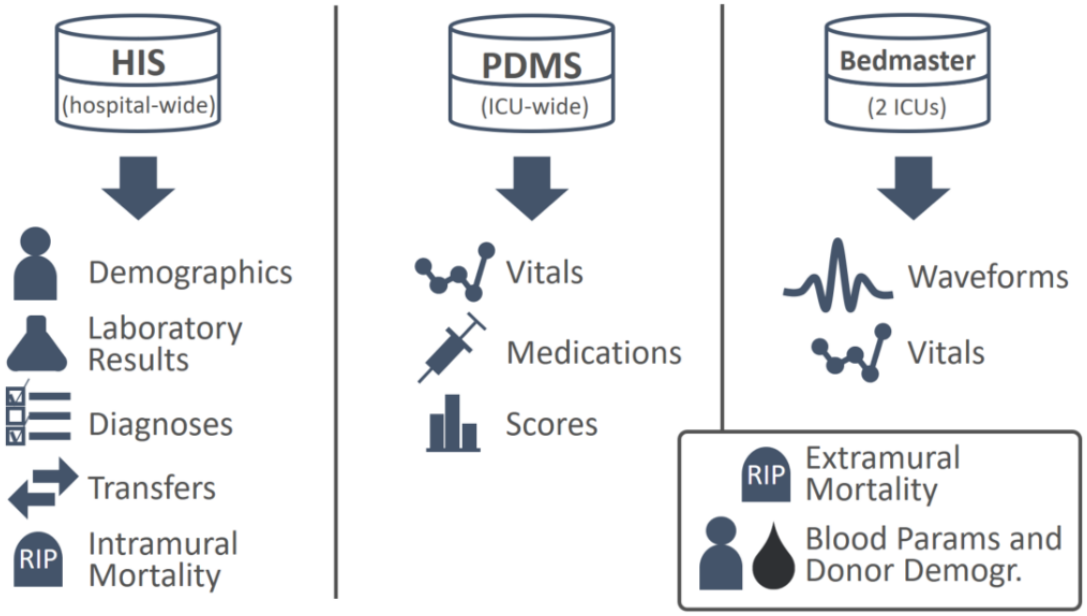
[c]ED: emergency department.

Data Sources

Relevant data for the 5 mentioned use cases are contained in 3 central data management systems in the hospital's IT infrastructure: the *hospital information system* (HIS), *patient data management system* (PDMS), and *Bedmaster system*. The HIS is a hospital-wide data warehouse that contains information on all patients admitted to the hospital. Among others, this includes demographics (date of birth, sex, etc), detailed information about in-hospital transfers, diagnoses, laboratory test results, and intramural mortality. PDMS is deployed in 5 ICUs associated with critical care in the hospital. Hence, it only contains information about patients admitted to the ICU during their hospital stay but complements the basic information found in the HIS with automatically recorded vital sign measurements (heart rate, blood pressure, body temperature, etc; up to 30 measurements per vital sign per hour), precise information about administered medications, and manually recorded scores (eg, Glasgow Coma Scale). The Bedmaster system [31] can be connected to bedside monitoring devices and automatically stores the vital signs, physiological waveforms, and alarms produced by these devices. The temporal resolution of the acquired data far exceeds the resolution in PDMS, with vitals being recorded every 2 seconds and waveforms sampled at rates of 60 to 240 Hz. This system is only deployed in 2 of the 5 ICUs and was installed in March 2018. Hence, the number of patients covered by it is significantly smaller compared with HIS and PDMS.

In addition, information about the extramural mortality of patients after hospital discharge was obtained from the Austrian Federal Statistics Agency (use case 1), and information about blood products transfused in the hospital was obtained from a local blood bank (use case 3). Figure 2 summarizes all data sources and modalities used in the 5 cases.

Figure 2. Data sources and exported modalities in use cases 1 to 5. HIS, PDMS, and Bedmaster are data management systems deployed in the hospital, whereas information about extramural mortality and blood products had to be obtained from external sources. HIS: hospital information system; PDMS: patient data management system; ICU: intensive care unit.



Ethics Approval

For each use case mentioned in this work, approval was obtained from the Ethics Committee of the Medical Faculty, Johannes Kepler University, Linz, Austria. The corresponding study numbers are 1015/2021, 1233/2020, 1232/2020, 1014/2021, and 1104/2020.

Results

Data Overview

Before beginning to export raw data from their hospital-internal storage, it is imperative to obtain an overview of what kind and how much data are available. This might sound obvious but can be more intricate than it seems. For example, the number of patients or cases at one's disposal is not always indicative of the amount of suitable data. Specifically, in the common setting of supervised learning, only data points to which clinically meaningful target labels can be assigned are useful. In the blood transfusion use case 3, for instance, the types of transfusion-related complications we could consider were limited by the availability of sufficient pre- and posttransfusion laboratory measurements to identify the respective complications. Ultimately, sufficient labeled training samples could only be generated for predicting acute kidney injury and acute respiratory failure. Other organ systems, although interesting in principle, had to be excluded from the analysis. Acute respiratory failure also had to be excluded eventually because the class imbalance was found to be too strong.

Given the richness of information stored in EHRs, there are normally enough data that can be converted into features that clinical prediction models may attend to. However, one must be aware that information accumulates over time, meaning that more data about a patient are available toward the end of the hospital or ICU stay than toward the beginning. For us, this was especially relevant in use case 2, where treatment priorities and 30-day mortality of patients visiting the ED had to be predicted based on only a few pieces of information typically recorded in the ED.

Defining Labels for Supervised Learning

Routinely collected retrospective EHR data do not always contain information about the outcomes that one wants to predict. Typical outcome parameters, apart from mortality or length of hospital stay, are often composites of several parameters that must be deduced from surrogate variables. Some authors, for instance, resort to hypotension as an indicator of cardiac instability [5,6], an approach we adopted in our use case 4 for predicting patient deterioration. Similarly, widely accepted criteria for organ system failure exist; for example, Kidney Disease: Improving Global Outcomes [32] for kidney disease and the Berlin definition for acute respiratory distress syndrome [33]. Both were used in use case 3.

Further problems can arise when trying to predict the effects of interventions. First, it might not always be possible to connect an observed outcome to a specific intervention, especially if multiple interventions occur within a short time. In the blood transfusion use case 3, in many cases, 2 or more blood products are transferred simultaneously, rendering it impossible to determine which of the administered transfusions are responsible for a posttransfusion complication. In such a situation, framing the prediction task as a *multiple instance learning problem* [34] might be the only remedy. Second, if the goal is to assess or improve existing clinical decision policies, one is confronted with questions such as "What would have happened to the patient if he/she had been treated differently?" Naturally, such questions are difficult to answer based on retrospective data in which interventions and treatments are fixed, and counterfactual trajectories cannot be explored, although the literature on estimating counterfactual treatment outcomes through statistical analysis and ML exists [35]. In use case 1, where the primary goal was to predict the optimal time for discharging ICU patients back to a ward, we resorted to answering the proxy question of whether transferred patients should have better stayed longer in the ICU. We determined this by identifying patients who died or returned unexpectedly shortly after ICU discharge.

Accessing and Extracting the Data

Hospital IT infrastructure is usually designed to provide easy access to the data of individual patients to deliver optimal care. Unfortunately, this does not imply that batches of data from distinct patients can be accessed, let alone extracted, easily. In particular, if the amount of manual interaction required for exporting data is too high, individual retrospective studies might be feasible, but the automatic real-time deployment of prediction models on live data may not be feasible. Data access can be challenging when there is only one source but even more so if there are multiple disparate data sources one must incorporate. In our project, we had to access 3 distinct databases: HIS, PDMS, and Bedmaster (Figure 2). HIS is a SAP-based system from which tables can be exported as CSV or Microsoft Excel files, and PDMS is a PostgreSQL relational database that allows exporting the results of queries in whatever table format is desired. In contrast, exporting data from the Bedmaster system turned out to be cumbersome because only XML and JSON exports are supported by default. Representing the massive amount of waveform and vital sign data in either of these verbose formats resulted in huge files that could not be processed efficiently; so, in the first step, we had to extract the relevant numerical values from the JSON files and store them in the more efficient HDF5 format. This process was considerably more intricate than anticipated because of inconsistencies in the exported data representation that are detailed in Multimedia Appendix 2 [36-39].

Matching Data From Different Sources

Data exported from different sources must be matched to obtain coherent records of the patients or cases under consideration. Under normal circumstances, this is straightforward because of common identifiers. However, according to our experience, such identifiers do not always need to be present or change over time. Specifically, data exported from the Bedmaster system lack identifiers, such as patient or case IDs. Knowing only the ICU bed they stem from, as well as the precise timestamp of each single recorded value, we had to assign the corresponding IDs manually based on the information about which patient occupied which ICU bed at which time. This approach works but is cumbersome and adds extra complexity and is another potential source of mistakes. It is also more difficult to automate than simply joining tables on common ID columns.

Mappings between identifiers and the entities they refer to may change over time as experienced in our project with drug codes. Every drug has a unique code that is used to reference it in prescription tables, but for unknown reasons, the coding changes at certain points in time. The precise information when this

happens is stored in another table, so that drug names *can* be recovered from the provided codes and timestamps of prescriptions. Yet again, the whole process is not as straightforward as we would have hoped.

Deidentification

Sensible personal information stored in EHRs can only be shared in a deidentified form. There are no universal rules *on how* data need to be deidentified, as long as identifying individual patients afterward becomes practically impossible. In our project, deidentification amounted to removing patient names and replacing hospital identifiers, such as patient IDs or case IDs, with project-internal identifiers that could be used to match corresponding data items across different tables. Furthermore, all timestamps were shifted by a random per-patient offset in the future to avoid reidentification of patients from knowing their exact admission or discharge times. Timestamps were shifted after matching data from different sources because some matching strategies depend on precise temporal information, as described earlier. The timestamps were shifted such that the time of day and day of week were preserved because both constitute potentially valuable information for downstream data analysis tasks. The same is true for seasonality, which was also roughly preserved. This deidentification policy is analogous to that used for MIMIC-III [21]. We remark that it is not as thorough as the policy implemented for releasing the more recent AmsterdamUMCdb [40]: there, theoretical concepts such as *k*-anonymity and *l*-diversity are considered to render reidentifying individual patients practically impossible under advanced threat models assuming “rogue researchers” and “rogue insurance companies” with access to the data. As, in our case, all data (even in deidentified form) are kept private and can only be accessed by project members, we did not deem such a thorough deidentification policy necessary.

Deidentification removes or replaces information that can otherwise be used to detect inconsistencies in the data, such as the same patient ID being accidentally assigned to multiple patients with different names. Therefore, it is crucial to ensure that any problems of this kind are detected and corrected either before or while deidentifying the data when the necessary information is still available. Specifically, we implemented extensive sanity checks that, for instance, ensure case and patient IDs are in a 1:n relationship (every case ID corresponds to a unique patient ID, but a patient ID can have multiple case IDs associated with it). All instances violating this principle are immediately reported to the human operator, allowing him or her to either overwrite one of the identifiers or discard the instances completely. Furthermore, missing patient IDs were automatically reconstructed from known case IDs whenever possible. The availability of patient IDs is essential because the random temporal offsets used for deidentifying timestamps are associated with patient IDs rather than case IDs. Finally, because hospital-assigned case IDs follow a clearly defined pattern that allows them to be distinguished from patient IDs, accidentally swapped case IDs and patient IDs are automatically exchanged before deidentification.

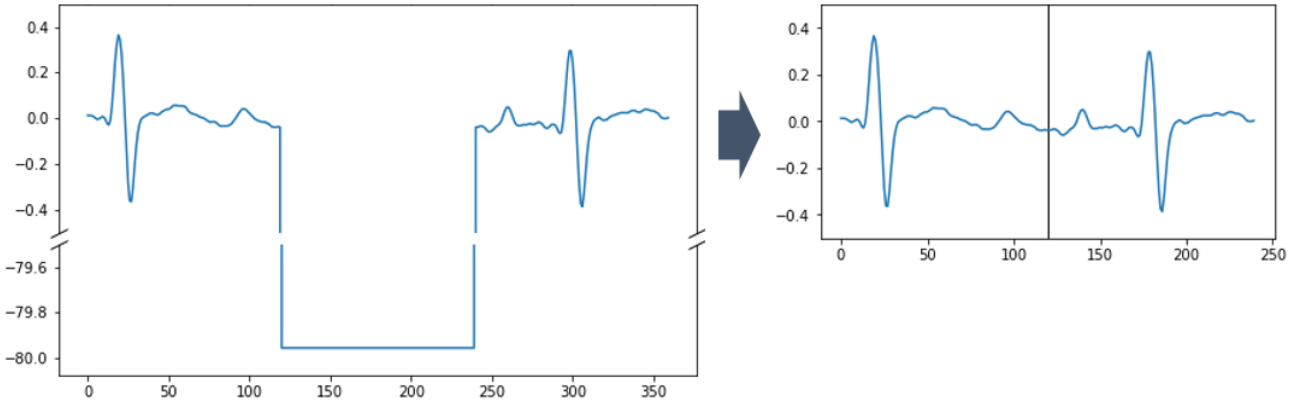
The kind of information that should be preserved by deidentification depends very much on the prediction task one wants to tackle. For example, in our approach, the temporal order of the data is preserved only within a patient but not across all patients. In particular, the total number of patients in the ICU at a certain point in time, a potentially relevant input for use case 1, can no longer be determined after deidentification. For the same reason, it is impossible to detect *domain shifts* in the deidentified data, which are systematic changes in the distribution of the data over time (domain shifts can be caused by many different factors such as new measurement equipment, laboratory test procedures, or changes in the prevalence of diseases in the patient population). Therefore, all relevant temporal features that could not be computed after deidentification had to be extracted and added to the data before deidentification.

Inspection and Exploratory Analysis

Real-world data can be corrupted or otherwise ill-behaved in many unforeseeable ways, in addition to well-known issues related to missing values or invalid measurements, that a thorough inspection and exploratory analysis is inevitable. Indeed, in our experience, this is one of the most labor-intensive tasks in the entire data preparation pipeline. Owing to the nature of the problem, it is difficult to devise general rules for what one should pay attention to. Instead, we report one particularly subtle issue encountered in our work. It might be specific to our hospital but is meant to serve as an illustrative example of what can unexpectedly happen when working with EHRs. More examples can be found in Multimedia Appendix 2.

In use case 4, we made heavy use of physiological waveform signals, such as electrocardiogram, arterial pressure, and oxygen saturation, to predict whether the condition of ICU patients will deteriorate within the next 15 minutes. Waveform signals are recorded by the Bedmaster system and can be exported as arrays of numerical values. It should be clear that because of the way in which these data are measured, there can be many types of measurement artifacts in the signals; that is, highly unusual waveform morphology caused by slipped sensors, or patient movements. This must be expected and addressed either explicitly by automatically detecting periods of invalid waveform data [36] or implicitly by relying on the subsequent ML algorithm’s ability to learn how to differentiate between normal and abnormal signals. An entirely unexpected issue is depicted in Figure 3: occasionally, the signals assume constant low values for a short time. The natural guess of measurement errors (eg, caused by slipped sensors) is likely wrong because simply cutting out the constant low-value period leads to smooth curves in all inspected cases. Such situations may thus indicate data artifacts of unknown origin that must be removed to obtain coherent signals, but strangely they do not always occur in all simultaneously recorded waveforms at the same time. Therefore, we opted to refrain from cutting out fragments of the raw signals to avoid a possible temporal misalignment of different waveforms.

Figure 3. Short periods of constant low values in waveform signals might have to be cut out. Left: original signal with a 0.5-second period of constant low values. Right: signal after cutting out the low value; as can be seen, the 2 ends of the signal fit perfectly.



Implementation

Eventually, the pipeline for extracting and preprocessing all relevant EHR data must be implemented in the actual code. This can be challenging in many respects. First, it is tempting to make extensive use of technologies aimed at rapid prototyping (eg, Jupyter notebooks) to quickly experiment with the data and preprocess them for a particular use case in a particular hard-coded way. This might work well in the short term; however, in the long term, a structured modular codebase that allows the exchange of individual components and adjustment (and logging!) of configuration settings is the better approach. In particular, the logging of configuration settings is of utmost importance to know exactly how data were preprocessed and how models were generated, thereby obtaining reproducible results.

Second, pipelines implemented to process a specific data modality for a particular use case should be reusable in other use cases that depend on the same data modality, at least to a certain extent. Even if the desired output format of the data after preprocessing differs between the 2 use cases, there are almost certainly some steps in the pipeline that are applicable to both. Reusing existing functionality rather than reimplementing it enables a consistent treatment of data across use cases and as a side effect may even help to abstract from the peculiarities of one use case and implement preprocessing functionality in a more general way. For example, we used laboratory test results in each of the 5 use cases either as features or for assigning labels (or both). In use case 4, the last 3 measurements of a fixed set of laboratory parameters relative to a given point in time are used as features, whereas in use case 3, the last measured value of a certain parameter before a blood transfusion is compared with measurements after the transfusion to determine whether it incurred a complication. Both are special cases of the more general principle of finding the last or first n measured values before or after a given point in time and could hence be implemented in one common function.

Finally, the inclusion of general-purpose third-party tools in the data preparation pipeline clearly has its benefits as well as potential downsides. On the one hand, functionality implemented therein does not have to be reimplemented (nor tested) from scratch, but on the other hand, Sculley et al [14] point out that it may lead to many glue code and pipeline jungles for bringing data into the right shape. In our project, we restricted ourselves to well-established libraries from the Python ecosystem, including NumPy [41], Pandas [42], and scikit-learn [43] and deliberately avoided tools such as FIDDLE [16]. The former 3 are libraries of useful classes and functions that can be easily integrated into one's own pipeline. The latter implements a full medical data preparation pipeline itself, which, although being generic and customizable in principle, did not offer the amount of flexibility we would have required to accommodate our data.

More precisely, our data preparation pipeline consists of 3 main steps: harmonization, validation, and transformation. Harmonization, that is, ensuring that equivalent concepts are represented consistently, is very specific to each data modality and typically amounts to assigning unique names to equivalent variables and converting measured values into a common unit of measurement. Validation of recorded values happens with respect to manually specified, threshold-based rules. Analogous to Harutyunyan et al [3], we distinguished between invalid numerical values and extreme outliers. Each validation rule is characterized by 2 ranges $r1 \subseteq r2$, where everything inside $r1$ is deemed admissible and everything outside $r2$ is deemed an extreme outlier. Extreme outliers $x \notin r2$ are deleted entirely, whereas the values $x \in r2 \setminus r1$ are set to the nearest admissible value in $r1$. Finally, transformation also depends on the specific data modality under consideration but is often concerned with resampling EHR tables in an event-based entity-attribute-value format into a more ML-friendly wide table format. This proceeds by aggregating all observations within a given time window with respect to a fixed set of rules, such as taking the mean, sum, or temporally last of all the measured values. If a variable has not been measured at all in a time window, the "missing" recordings are imputed. As other authors have noted [44], clinical measurements are not missing at random; therefore, explicit *missingness masks* indicating whether a value has been imputed are added as extra features. In general, one must also be careful when imputing the mean or median of all observed values, as this could introduce bias. For example, if a variable is only measured if a patient has a certain condition, the measured values are not representative of the entire population.

Discussion

Principal Findings

The preceding sections illustrate that the preparation of EHRs for secondary analysis and the development of prediction models constitute a challenging endeavor. In addition to the well-known ubiquitous data problems for which generic off-the-shelf solutions exist (eg, imputation of missing values), we identified many issues in our raw data that had to be addressed individually. Even worse, none of these issues could be expected or popped up during the first quick scan of the data but instead were discovered only after a thorough exploratory analysis. Different kinds of patient identifiers being accidentally swapped is certainly something one would not expect at all, yet we found a few such cases in our data. The use of multiple codes or names for the same clinical concept is also not trivial to detect, especially if it is a mere artifact of the internal data representation that does not surface in clinical practice. If the mapping between codes and concepts changes over time, data harmonization becomes a true challenge. With regard to data validation, blindly discarding all nonnumeric values of a supposedly numeric variable fails to account for censored values such as ">120.0" (Multimedia Appendix 2) that do carry useful information. Finally, the subtle issues with waveform data reported above not only demand a thorough systematic analysis of timestamps and measured values but are also difficult to fix. Altogether, these observations support our claim that although generic tools such as FIDDLE [16] and Clairvoyance [25] doubtlessly do have their merits, one must be careful not to underestimate the additional effort of modality- and source-specific data analysis and preparation. In general, we believe that extensive libraries of well-documented, generic, and cleanly implemented functionalities focusing on the peculiarities of medical data preparation (harmonizing and validating physiological variables, resampling event-based entity-attribute-value tables into wide tables, etc) are more valuable than full-fledged end-to-end pipelines, regardless of how generic and configurable they are.

Extracting labels that indicate the outcome of interest from retrospective data can be more intricate than one might expect. Often, these outcomes (patient deterioration, organ system failure, optimal treatment policy, etc) are not explicitly recorded in EHRs and must therefore be approximated. The quality of such an approximation might influence not only the performance of the generated prediction models but also their applicability to clinical practice. Furthermore, if the definition of some label depends on scarcely recorded variables, only a few labeled samples may remain. In such a situation, methods based on *self-supervised and semisupervised learning* [45-47] might be the only remedy.

EHRs contain highly sensitive patient information that, for good reasons, must be deidentified before it can be shared with scientific partners in research projects. How and to what extent this needs to be carried out often not clearly defined, especially regarding the treatment of temporal information. Temporal data may contain highly relevant information depending on the concrete use case. On the one hand, knowing the time of day and day of week of a particular event is necessary if the prediction task at hand has to take clinical routines into account; on the other hand, knowing the (rough) order of events across different patients enables detecting domain shifts in the underlying data distribution. Finally, if the use of a particular resource at any given point in time is of interest, this information must be extracted before deidentifying the timestamps, or timestamp deidentification must be avoided entirely. In our experience, it is good to first determine the kind of information one needs for a particular use case and then devise deidentification strategies that preserve as much of the previously determined information as possible while observing legal regulations and hospital-internal restrictions.

Finally, if the ultimate goal of developing prediction models is to deploy them in clinical practice, data access becomes a factor that must be considered. The more manual steps involved in exporting the data from the hospital IT infrastructure into the desired format, the more difficult real-time deployment will be. In our use case 4, automatically exporting the necessary data of all current ICU patients after every n minutes and then promptly processing them is challenging and currently work in progress. This mainly owes to the fact that the entire data warehousing system of the Kepler University Hospital was designed for clinical use rather than real-time analysis. However, alternatives exist; a sophisticated solution for efficient storage of and access to medical data for data science projects is presented in a study by McPadden et al [48].

Workflow

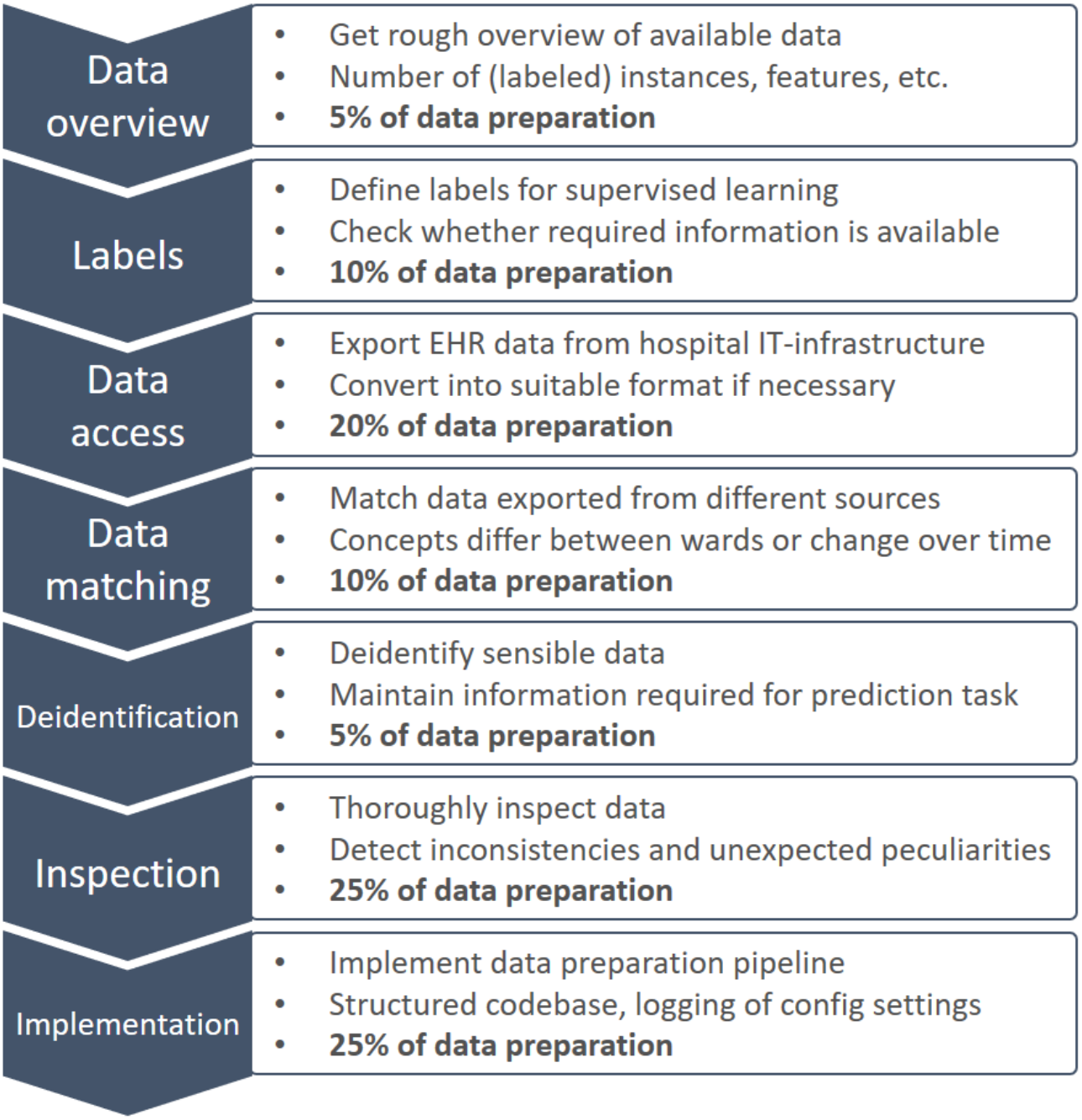
The data preparation workflow we followed in our project is summarized in Figure 4, with rough estimates of the relative time and effort taken by the individual steps. We think that it generalizes to other data science projects with retrospective EHR data and hope that it can serve as guidance for other researchers to identify and address potential problems early and avoid some common pitfalls.

The presentation of the (linear) workflow in Figure 4 is simplified because in reality, there are many feedback loops. For instance, inspecting the data may reveal issues that can only be rectified if additional information is extracted from the system, and some issues might only surface after developing the first prediction models.

It is important to note that the results presented in this paper only refer to data preparation for subsequent model development but not to the development and validation of actual prediction models. We think these are "standard" tasks in data science and ML that are not specific to medical data. However, we do acknowledge that selecting the appropriate class of prediction models for a given task, optimizing hyperparameters, and training models in the right way are by no means trivial and require a lot of time and effort. This is also true for deploying models in clinical practice, where topics such as *handling domain shifts*, *detecting out-of-distribution data*, and *explaining model decisions in a manner comprehensible to patients* must be addressed. Things become even more difficult if existing models are to be deployed in other hospitals because most of the steps in the above workflow must be repeated. Only the definition of labels and (possibly) deidentification can be

skipped, and some parts of the existing pipeline implementation can perhaps be reused. According to our rough estimate, approximately 75% of the effort invested in the initial data preparation for developing prediction models must be reinvested for each hospital that these models are deployed. As noted in a study by Sendak et al [30], this incurs significant additional costs.

Figure 4. Data preparation workflow for retrospective EHR data analysis. EHR: electronic health record.



Conclusions

Preparing raw medical data from productive environments for retrospective analysis and ML remains challenging and time consuming. Our findings suggest that real-world EHR data can be messy and corrupted in so many subtle ways that thorough exploratory analysis and tailor-made preprocessing functionality for the data at hand are inevitable. We want to create awareness of this fact and hope that the sketched data preparation workflow becomes a valuable guidance for future large-scale data science projects involving routinely acquired medical data.

Acknowledgments

This research was funded by Medical Cognitive Computing Center (MC^3) and supported by the strategic economic research program “Innovatives OÖ 2020” of the province of Upper Austria. This project was cofinanced by research subsidies granted by

the government of Upper Austria and supported by the University SAL Labs initiative of Silicon Austria Labs (SAL) and its Austrian partner universities for applied fundamental research for electronic based systems. RISC Software GmbH is a member of the Upper Austrian Research Innovation Network. The authors thank Philipp Moser for carefully proofreading a draft version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Brief evaluation and comparison of four EHR data preparation tools.
[DOCX File , 18 KB - medinform_v10i10e38557_app1.docx]

Multimedia Appendix 2
Overview of the data modalities used in our research and specific issues encountered when processing them.
[DOCX File , 137 KB - medinform_v10i10e38557_app2.docx]

References

1. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med 2018 May 8;1:18 [FREE Full text] [doi: 10.1038/s41746-018-0029-1] [Medline: 31304302]
2. Purushotham S, Meng C, Che Z, Liu Y. Benchmarking deep learning models on large healthcare datasets. J Biomed Inform 2018 Jul;83:112-134 [FREE Full text] [doi: 10.1016/j.jbi.2018.04.007] [Medline: 29879470]
3. Harutyunyan H, Khachatrian H, Kale DC, Ver Steeg G, Galstyan A. Multitask learning and benchmarking with clinical time series data. Sci Data 2019 Jun 17;6(1):96 [FREE Full text] [doi: 10.1038/s41597-019-0103-9] [Medline: 31209213]
4. Caicedo-Torres W, Gutierrez J. ISeeU: visually interpretable deep learning for mortality prediction inside the ICU. J Biomed Inform 2019 Oct;98:103269 [FREE Full text] [doi: 10.1016/j.jbi.2019.103269] [Medline: 31430550]
5. Hatib F, Jian Z, Buddi S, Lee C, Settels J, Sibert K, et al. Machine-learning algorithm to predict hypotension based on high-fidelity arterial pressure waveform analysis. Anesthesiology 2018 Oct;129(4):663-674 [FREE Full text] [doi: 10.1097/ALN.0000000000002300] [Medline: 29894315]
6. Hyland SL, Faltys M, Hüser M, Lyu X, Gumbsch T, Esteban C, et al. Early prediction of circulatory failure in the intensive care unit using machine learning. Nat Med 2020 Mar;26(3):364-373. [doi: 10.1038/s41591-020-0789-4] [Medline: 32152583]
7. Corey KM, Kashyap S, Lorenzi E, Lagoo-Deenadayalan SA, Heller K, Whalen K, et al. Development and validation of machine learning models to identify high-risk surgical patients using automatically curated electronic health record data (Pythia): a retrospective, single-site study. PLoS Med 2018 Nov 27;15(11):e1002701 [FREE Full text] [doi: 10.1371/journal.pmed.1002701] [Medline: 30481172]
8. Bottino F, Tagliente E, Pasquini L, Napoli AD, Lucignani M, Figà-Talamanca L, et al. COVID mortality prediction with machine learning methods: a systematic review and critical appraisal. J Pers Med 2021 Sep 07;11(9):893 [FREE Full text] [doi: 10.3390/jpm11090893] [Medline: 34575670]
9. Tschoellitsch T, Dünser M, Böck C, Schwarzbauer K, Meier J. Machine learning prediction of SARS-CoV-2 polymerase chain reaction results with routine blood tests. Lab Med 2021 Mar 15;52(2):146-149 [FREE Full text] [doi: 10.1093/labmed/lmaa111] [Medline: 33340312]
10. Yang D, Martinez C, Visuña L, Khandhar H, Bhatt C, Carretero J. Detection and analysis of COVID-19 in medical images using deep learning techniques. Sci Rep 2021 Oct 04;11(1):19638 [FREE Full text] [doi: 10.1038/s41598-021-99015-3] [Medline: 34608186]
11. Roland T, Böck C, Tschoellitsch T, Maletzky A, Hochreiter S, Meier J, et al. Domain shifts in machine learning based Covid-19 diagnosis from blood tests. J Med Syst 2022 Mar 29;46(5):23 [FREE Full text] [doi: 10.1007/s10916-022-01807-1] [Medline: 35348909]
12. Guleria P, Ahmed S, Alhumam A, Srinivasu PN. Empirical study on classifiers for earlier prediction of COVID-19 infection cure and death rate in the Indian states. Healthcare (Basel) 2022 Jan 02;10(1):85 [FREE Full text] [doi: 10.3390/healthcare10010085] [Medline: 35052249]
13. 2016 Data Science Report. CrowdFlower. 2016. URL: https://visit.figure-eight.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf; [accessed 2022-09-23]
14. Sculley D, Holt G, Golovin D, Davydov E, Phillips T, Ebner D, et al. Hidden technical debt in machine learning systems. In: Proceedings of the 2015 Advances in Neural Information Processing Systems. 2015 Presented at: NeurIPS '15; December 7-12, 2015; Montreal, Canada p. 2503-2511.
15. Wang S, McDermott MB, Chauhan G, Ghassemi M, Hughes MC, Naumann T. MIMIC-Extract: a data extraction, preprocessing, and representation pipeline for MIMIC-III. In: Proceedings of the ACM Conference on Health, Inference, and Learning. 2020 Apr Presented at: CHIL '20; April 2-4, 2020; Toronto, Canada p. 222-235. [doi: 10.1145/3368555.3384469]

16. Tang S, Davarmanesh P, Song Y, Koutra D, Sjoding MW, Wiens J. Democratizing EHR analyses with FIDDLE: a flexible data-driven preprocessing pipeline for structured clinical data. J Am Med Inform Assoc 2020 Dec 09;27(12):1921-1934 [FREE Full text] [doi: 10.1093/jamia/ocaa139] [Medline: 33040151]
17. Mandyam A, Yoo EC, Soules J, Laudanski K, Engelhardt BE. COP-E-CAT: cleaning and organization pipeline for EHR computational and analytic tasks. In: Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics. 2021 Aug Presented at: BCB '21; August 1-4, 2021; Gainesville, FL, USA p. 1-9. [doi: 10.1145/3459930.3469536]
18. Johnson AE, Ghassemi MM, Nemati S, Niehaus KE, Clifton DA, Clifford GD. Machine learning and decision support in critical care. Proc IEEE Inst Electr Electron Eng 2016 Feb;104(2):444-466 [FREE Full text] [doi: 10.1109/JPROC.2015.2501978] [Medline: 27765959]
19. Attribution 4.0 International (CC BY 4.0). Creative Commons. URL: https://creativecommons.org/licenses/by/4.0/ [accessed 2022-10-05]
20. Mandel JC, Kreda DA, Mandl KD, Kohane IS, Ramoni RB. SMART on FHIR: a standards-based, interoperable apps platform for electronic health records. J Am Med Inform Assoc 2016 Sep;23(5):899-908 [FREE Full text] [doi: 10.1093/jamia/ocv189] [Medline: 26911829]
21. Johnson AE, Pollard TJ, Shen L, Lehman LW, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016 May 24;3:160035 [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
22. Johnson AE, Stone DJ, Celi LA, Pollard TJ. The MIMIC code repository: enabling reproducibility in critical care research. J Am Med Inform Assoc 2018 Jan 01;25(1):32-39 [FREE Full text] [doi: 10.1093/jamia/ocx084] [Medline: 29036464]
23. Johnson A, Bulgarelli L, Pollard T, Horng S, Celi LA, Mark R. MIMIC-IV (version 2.0). PhysioNet 2022. [doi: 10.13026/7vcr-e114]
24. Pollard TJ, Johnson AE, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU collaborative research database, a freely available multi-center database for critical care research. Sci Data 2018 Sep 11;5:180178 [FREE Full text] [doi: 10.1038/sdata.2018.178] [Medline: 30204154]
25. Jarrett D, Yoon J, Bica I, Qian Z, Ercole A, van der Schaar M. Clairvoyance: a pipeline toolkit for medical time series. In: Proceedings of the 8th International Conference on Learning Representations. 2020 Presented at: ICLR '20; April 26-30, 2020; Addis Ababa, Ethiopia p. 1-16.
26. Shi X, Prins C, Van Pottelbergh G, Mamouris P, Vaes B, De Moor B. An automated data cleaning method for Electronic Health Records by incorporating clinical knowledge. BMC Med Inform Decis Mak 2021 Sep 17;21(1):267 [FREE Full text] [doi: 10.1186/s12911-021-01630-7] [Medline: 34535146]
27. Weiskopf NG, Bakken S, Hripcsak G, Weng C. A data quality assessment guideline for electronic health record data reuse. EGEMS (Wash DC) 2017 Sep 04;5(1):14 [FREE Full text] [doi: 10.5334/egems.218] [Medline: 29881734]
28. Mashoufi M, Ayatollahi H, Khorasani-Zavareh D. A review of data quality assessment in emergency medical services. Open Med Inform J 2018 May 31;12:19-32 [FREE Full text] [doi: 10.2174/1874431101812010019] [Medline: 29997708]
29. Terry AL, Stewart M, Cejic S, Marshall JN, de Lusignan S, Chesworth BM, et al. A basic model for assessing primary health care electronic medical record data quality. BMC Med Inform Decis Mak 2019 Feb 12;19(1):30 [FREE Full text] [doi: 10.1186/s12911-019-0740-0] [Medline: 30755205]
30. Sendak MP, Balu S, Schulman KA. Barriers to achieving economies of scale in analysis of EHR data. A cautionary tale. Appl Clin Inform 2017 Aug 09;8(3):826-831 [FREE Full text] [doi: 10.4338/ACI-2017-03-CR-0046] [Medline: 28837212]
31. BedMasterEx. Data acquisiton and infinite storage of medical device data. Anandic Medical Systems. URL: https://www.bedmaster.net/en/products/bedmasterex; [accessed 2022-09-23]
32. Ostermann M, Bellomo R, Burdmann EA, Doi K, Endre ZH, Goldstein SL, Conference Participants. Controversies in acute kidney injury: conclusions from a Kidney Disease: Improving Global Outcomes (KDIGO) Conference. Kidney Int 2020 Aug;98(2):294-309 [FREE Full text] [doi: 10.1016/j.kint.2020.04.020] [Medline: 32709292]
33. ARDS Definition Task Force, Ranieri VM, Rubenfeld GD, Thompson BT, Ferguson ND, Caldwell E, et al. Acute respiratory distress syndrome: the Berlin definition. JAMA 2012 Jun 20;307(23):2526-2533. [doi: 10.1001/jama.2012.5669] [Medline: 22797452]
34. Cheplygina V, Tax DM, Loog M. Multiple instance learning with bag dissimilarities. Pattern Recognit 2015 Jan;48(1):264-275. [doi: 10.1016/j.patcog.2014.07.022]
35. Bica I, Alaa AM, Jordon J, van der Schaar M. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. In: Proceedings of the 8th International Conference on Learning Representations. 2020 Presented at: ICLR '20; April 26-30, 2020; Addis Ababa, Ethiopia p. 1-26. [doi: 10.48550/arXiv.2002.04083]
36. De Ryck T, De Vos M, Bertrand A. Change point detection in time series data using autoencoders with a time-invariant representation. IEEE Trans Signal Process 2021;69:3513-3524. [doi: 10.1109/TSP.2021.3087031]
37. Phan HT, Borca F, Cable D, Batchelor J, Davies JH, Ennis S. Automated data cleaning of paediatric anthropometric data from longitudinal electronic health records: protocol and application to a large patient cohort. Sci Rep 2020 Jun 23;10(1):10164 [FREE Full text] [doi: 10.1038/s41598-020-66925-7] [Medline: 32576940]
38. International Statistical Classification of Diseases and Related Health Problems. 10th Revision. World Health Organization. 2019. URL: https://icd.who.int/browse10/2019/en; [accessed 2022-09-23]

39. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997 Nov 15;9(8):1735-1780. [doi: 10.1162/neco.1997.9.8.1735] [Medline: 9377276]
40. Thoral PJ, Peppink JM, Driessen RH, Sijbrands EJ, Kompanje EJ, Kaplan L, Amsterdam University Medical Centers Database (AmsterdamUMCdb) Collaborators and the SCCM/ESICM Joint Data Science Task Force. Sharing ICU patient data responsibly under the society of critical care medicine/European Society of Intensive Care Medicine Joint Data Science Collaboration: the Amsterdam University Medical Centers Database (AmsterdamUMCdb) example. Crit Care Med 2021 Jun 01;49(6):e563-e577 [FREE Full text] [doi: 10.1097/CCM.0000000000004916] [Medline: 33625129]
41. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. Nature 2020 Sep;585(7825):357-362 [FREE Full text] [doi: 10.1038/s41586-020-2649-2] [Medline: 32939066]
42. McKinney W. Data structures for statistical computing in Python. In: Proceedings of the 9th Python in Science Conference. 2010 Presented at: SciPy '10; June 28-July 3, 2010; Austin, TX, USA p. 56-61. [doi: 10.25080/Majora-92bf1922-00a]
43. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res 2011;12(2011):2825-2830.
44. Li J, Yan XS, Chaudhary D, Avula V, Mudiganti S, Husby H, et al. Imputation of missing values for electronic health record laboratory data. NPJ Digit Med 2021 Oct 11;4(1):147 [FREE Full text] [doi: 10.1038/s41746-021-00518-0] [Medline: 34635760]
45. Vincent P, Larochelle H, Bengio Y, Manzagol PA. Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th International Conference on Machine Learning. 2008 Presented at: ICML '08; July 5-9, 2008; Helsinki, Finland p. 1096-1103. [doi: 10.1145/1390156.1390294]
46. Yoo J, Zhang Y, Jordon J, van der Schaar M. VIME: extending the success of self- and semi-supervised learning to tabular domain. In: Proceedings of the 2020 Advances in Neural Information Processing Systems. 2020 Presented at: NeurIPS '20; December 6-12, 2020; Virtual p. 11033-11043.
47. Arik S, Pfister T. TabNet: attentive interpretable tabular learning. Proc AAAI Conf Artif Intell 2021 May 18;35(8):6679-6687. [doi: 10.1609/aaai.v35i8.16826]
48. McPadden J, Durant TJ, Bunch DR, Coppi A, Price N, Rodgerson K, et al. Health care and precision medicine research: analysis of a scalable data science platform. J Med Internet Res 2019 Apr 09;21(4):e13043 [FREE Full text] [doi: 10.2196/13043] [Medline: 30964441]

Abbreviations

ED: emergency department
EHR: electronic health record
eICU-CRD: telehealth ICU collaborative research database
HIS: hospital information system
ICU: intensive care unit
MIMIC: Medical Information Mart for Intensive Care
ML: machine learning
PDMS: patient data management system

Edited by C Lovis; submitted 07.04.22; peer-reviewed by L Celi, FM Calisto, M Sendak; comments to author 09.07.22; revised version received 02.08.22; accepted 07.09.22; published 21.10.22.

Please cite as:
Maletzky A, Böck C, Tschoellitsch T, Roland T, Ludwig H, Thumfart S, Giretzlehner M, Hochreiter S, Meier J
Lifting Hospital Electronic Health Record Data Treasures: Challenges and Opportunities
JMIR Med Inform 2022;10(10):e38557
URL: https://medinform.jmir.org/2022/10/e38557
doi:10.2196/38557
PMID:36269654

Original Paper

Appropriateness of Alerts and Physicians' Responses With a Medication-Related Clinical Decision Support System: Retrospective Observational Study

Hyunjung Park[1*], BA; Minjung Kathy Chae[1], MD, MSc; Woohyeon Jeong[1], BA; Jaeyong Yu[1], MS; Weon Jung[1], BA; Hansol Chang[1,2], MD; Won Chul Cha[1,2,3*], MD, PhD

[1]Department of Digital Health, Samsung Advanced Institute of Health Sciences & Technology, Sungkyunkwan University, Seoul, Republic of Korea
[2]Department of Emergency Medicine, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea
[3]Digital Innovation Center, Samsung Medical Center, Seoul, Republic of Korea
[*]these authors contributed equally

Corresponding Author:
Won Chul Cha, MD, PhD
Department of Digital Health
Samsung Advanced Institute of Health Sciences & Technology
Sungkyunkwan University
81 Irwon-ro
Gangnam-gu
Seoul
Republic of Korea
Phone: 82 2 3410 2053
Fax: 82 2 3410 0012
Email: docchaster@gmail.com

Abstract

Background: Alert fatigue is unavoidable when many irrelevant alerts are generated in response to a small number of useful alerts. It is necessary to increase the effectiveness of the clinical decision support system (CDSS) by understanding physicians' responses.

Objective: This study aimed to understand the CDSS and physicians' behavior by evaluating the clinical appropriateness of alerts and the corresponding physicians' responses in a medication-related passive alert system.

Methods: Data on medication-related orders, alerts, and patients' electronic medical records were analyzed. The analyzed data were generated between August 2019 and June 2020 while the patient was in the emergency department. We evaluated the appropriateness of alerts and physicians' responses for a subset of 382 alert cases and classified them.

Results: Of the 382 alert cases, only 7.3% (n=28) of the alerts were clinically appropriate. Regarding the appropriateness of the physicians' responses about the alerts, 92.4% (n=353) were deemed appropriate. In the classification of alerts, only 3.4% (n=13) of alerts were successfully triggered, and 2.1% (n=8) were inappropriate in both alert clinical relevance and physician's response. In this study, the override rate was 92.9% (n=355).

Conclusions: We evaluated the appropriateness of alerts and physicians' responses through a detailed medical record review of the medication-related passive alert system. An excessive number of unnecessary alerts are generated, because the algorithm operates as a rule base without reflecting the individual condition of the patient. It is important to maximize the value of the CDSS by comprehending physicians' responses.

(JMIR Med Inform 2022;10(10):e40511) doi:10.2196/40511

KEYWORDS

clinical decision support system; computerized physician order entry; alert fatigue; health personnel; decision-making support; physician behavior; physician response; alert system

Introduction

Background

Computerized physician order entry (CPOE), linked to a clinical decision support system (CDSS), has become essential in the health care system. The main purpose of a CDSS is to improve patient safety and quality of care, and a medication-related CDSS is especially valuable [1,2]. In a medication-related CDSS, the alerting system provides dosing guidance or drug-drug, drug-allergy, and drug-age warnings that help clinicians prescribe correct orders. Early studies on CDSSs prompted substantial anticipation that medication-related CDSSs, such as alerting systems, may prevent adverse events and enhance patient safety [3,4].

Despite the increasing implementation of CDSS alerts, a substantial number of alerts are overridden [5-7]. The alert override rate is high, sometimes up to 96% [5]. Override is often invoked for reasons such as low alert specificity (ie, a lack of clinical relevance) and inadequate alert content [8,9]. Low alert acceptance was associated with repeated alerts that are inappropriate [6,10]. Excessive alerts that are not clinically relevant could lead to alert fatigue and contribute to alert overrides [11,12].

A common issue connected with the implementation of clinical decision support tools in electronic medical records (EMRs) is alert fatigue [13]. Alert fatigue is the issue in which users of a CDSS that generates an excessive amount of warning messages tend to overlook the majority of these alerts, including those that warn them of potentially clinically relevant errors [2]. A CDSS can fail to enhance patient safety due to alert fatigue. Alert fatigue arises when an excessive number of irrelevant alerts drives users to routinely override them [14].

In the CDSS, 2 types of alerts are usually used. One type of alerts is active or “pop-up” warnings. These alerts require an action from the user for the clinical process to continue, such as clicking a button or stating the overriding reason. The other type of alerts is passive warnings, such as flagging potentially abnormal values. Passive alerts, unlike active alerts, do not interrupt the provider’s workflow; hence, these alerts do not require a response from the user to override the clinical process. Numerous studies have established the issue of alert fatigue with active alerts [10,12,15,16]. Passive alerts may also be a substantial cause of alert fatigue. The true burden of these alerts has rarely been assessed [17].

There is limited research evaluating the appropriateness of overrides with no override reasons in the passive alert system and the alert itself for clinical appropriateness for a patient’s specific condition. To understand the behavior of physicians, previous studies have only evaluated the appropriateness of overrides based on their reasoning [1,18]. In this study, we evaluated the appropriateness of alerts and physicians’ responses in a passive alert system through a patient EMR. We also categorized the alerts assessed by clinical relevance and physicians’ responses. This study may provide insights into the clinical use of medication alerts, whether physicians override them, and what reactions physicians offer when responding to them.

Objective

This study aimed to evaluate the clinical appropriateness of alerts and the corresponding physicians’ responses in a medication-related passive alert system.

Methods

Study Design

This study was a retrospective observational study with stratified sampling according to medication. The analyzed alerts were generated from medication orders between August 2019 and June 2020 in the emergency department (ED). We obtained medication orders, alerts, and patient EMR data from a clinical data warehouse (CDW). In Korea, it is stipulated by law that only physicians can prescribe orders, except in a limited number of cases.

Ethics Approval

This study was approved by the Institutional Review Board of the Samsung Medical Center (IRB 2021-09-115).

Study Setting

This study was conducted in the ED of a tertiary academic medical center in Seoul, Korea. It serves 2 million outpatient visits annually and provides in-hospital service for 1975 beds. The ED has 69 beds and approximately 35 doctors. The annual number of patients visiting the ED ranges from 75,000 to 80,000. The workflow of the ED is uncontrolled and unpredictable [19]. Adverse events following an ED visit were reported less frequently but were more preventable than in other hospital settings [20]. Since the ED has various medication prescription patterns, diverse alerts can be analyzed by checking the patients in the ED.

EMR System and Medication Order (Prescription) System

Our EMR system is a self-developed system implemented in 2016. Data Analytics and Research Window for Integrated Knowledge (DARWIN) is an extensive system that includes CPOE as well as nursing, pharmacy, billing, and research support and even patient portal and web services.

CDSS Design: Passive Alert System

A passive alert system in the medication CDSS was applied to the DARWIN. Although passive alerts with in-line text do not interfere with physicians’ workflow, they may also result in decreased effectiveness of the CDSS alerts [21]. The alert appears before the order is confirmed. A response is not required to allow the prescription. The rule-based database for the CDSS was supplied by the KIMS POC knowledge base (KIMS Co) with weekly updates. The types of alerts were age, allergy, dose, drug-drug interaction (DDI), and renal.

CDW Use

This study was performed using data extracted from the CDW at the study site. The CDW is an integrated storage for clinical data that are updated daily, such as deidentified patient

demographic information, diagnosis, prescription, and laboratory results. In the past, researchers had to check the variables required for research individually and process the data accordingly. However, using the CDW, researchers can easily obtain the data automatically, sorted according to the various variables assumed by the researcher. CDW supports the automatic conversion of unstructured data, such as text to standardized data, to make it possible to conduct prospective cohort studies conveniently.

Selection of Alerts

In all, 20 frequently overridden medication alerts were selected. We thought that alerts that are frequently overridden would be less clinically relevant; therefore, we prioritized alerts that are frequently overridden as evaluation targets. DDI types and alerts that are difficult to evaluate for clinical appropriateness were excluded as follows: when there was no specific dose setting information for reduction and when the range of dose adjustment according to the indication and severity was wide. Overridden cases and nonoverridden cases were randomly extracted from 20 frequently overridden medication alerts. The number of cases for each medication alert are shown below.

Definition of Alert Overrides and Appropriateness

Alert overrides occur when physicians do not change orders as suggested by the alert. Our previous study defined an alert override as no change in order when an alert occurred on the log data [22]. In this study, however, alert override means no change in order when an alert occurred or a re-order of the same prescription later. In nonoverridden cases, many physicians prescribed the nonoverridden order again, and we considered this case to be an override. If the identical prescription that generated the alert was given to the same patient within 48 hours, it was deemed an override. Alert clinical relevance means that the alert is suitable for each patient's condition and that the alert actually helped the physician order the prescription. The physicians' response appropriateness indicates whether the physicians' override or nonoverride was appropriate considering the patient's clinical condition.

Detailed Medical Record Review

Through advanced medical record reviews of alert overridden cases and literature research, a group of 3 clinicians (a physician, a pharmacist, and a nurse) determined the criteria for the appropriateness of each alert. In a detailed medical record review, information such as the patient's age, gender, weight, laboratory results (potassium, sodium, serum creatinine, or glomerular filtration rate, etc), and computed tomography status was confirmed through the patient's EMR. Each group member independently reviewed random samples of the 382 alert cases for the evaluation of the appropriateness of alert clinical relevance and physicians' responses. When panel members disagreed, consensus was reached via group discussion.

Classification of Alerts

The alerts were classified based on the results of the appropriateness evaluation. We referred to the evaluation framework developed by McCoy et al [23]. Since the passive alert system does not collect the overriding reason, it may be difficult to judge the appropriateness. Therefore, we included a nondecidable category in the alert classification table (Figure 1).

Figure 1. Classification table for alerts. The alert classification table included the nondecidable category—since the passive alert system does not include an override reason, some cases might be difficult to evaluated.

		Physicians' response		
		Appropriate	Inappropriate	Nondecidable
Alert clinical relevance	Appropriate	Successful alerts	Physicians' nonadherence	Response nondecidable
	Inappropriate	Justifiable overrides	Unintended adverse consequences	Response nondecidable
	Nondecidable	Alert nondecidable	Alert nondecidable	nondecidable

Korean Triage and Acuity Scale (KTAS)

The KTAS is an evaluation tool used to categorize the severity and urgency of ED patients. It is a 5-level triage scale based on the severity of the patient's chief complaint and symptoms. The KTAS was established in 2012 in Korea in an effort to enhance patient safety and minimize ED congestion at the hospital level. Patients who enter the ED are evaluated by KTAS using the following procedure: impression evaluation, infection confirmation, primary symptom selection, and primary/secondary considerations [24,25].

Data Analysis

Commonly overridden medications were subgrouped according to alert type, and alert patterns were examined. Samples for the medical record review were extracted using stratified random sampling. In our samples, we analyzed the appropriateness of alerts, physicians' responses, and patient demographics. Interrater reliability for the evaluation of alert and physicians'

response appropriateness was calculated by using a κ index. The results are presented as counts and percentages. The rate of false positive alerts, physicians' response inappropriateness, and override were expressed as percentages of total alerts. All statistical tests were performed using R statistical software (version 4.0.3; R Foundation for Statistical Computing).

Results

Figure 2 shows the detailed selection process for medication alert data. A total of 39,286 (10.5% alert rate) CDSS alerts occurred for 374,133 medication orders between August 2019 and June 2020. We selected 20 frequently overridden medication alerts stratified by the medication alert type (Table 1). The number of alert cases analyzed for medical record reviews was 382 (200 overridden and 182 nonoverridden cases).

The medical record review included 356 patients. Table 2 shows the demographic information of the patients in the medical record review cases. Overall, the patients' basic characteristics showed that the majority were men (204/356, 57.3%), aged more than 60 years (205/356, 57.6%), and had KTAS scores of 3 (197/356, 55.3%).

A total of 728 medications triggered an alarm; however, we chose 20 frequently overridden medication alerts, because we thought that alerts that are frequently overridden would be less clinically relevant. Table 1 shows the 20 analyzed medications. In the overridden case, all medication alerts included 10 cases; however, in the nonoverridden case, methylprednisolone (n=6), epinephrine (n=9), cefditoren (n=2), cefazolin (n=6), and ampicillin/sulbactam (n=9) had fewer than 10 cases.

Table 3 shows the results of the appropriateness evaluation for alert clinical relevance and physicians' responses. Interestingly, of the 382 alert cases, the only 7.3% (n=28) were clinically relevant alerts. In the physicians' response assessment, 92.4% (n=353) were appropriate and 1.6% (n=6) were nondecidable. The interrater reliability for alert clinical relevance appropriateness and physicians' response appropriateness were moderate (κ=0.47) and fair (κ=0.28), respectively. In our study, there was no difference in the appropriateness of clinical relevance between overridden and nonoverridden alerts. When an overridden alert and a nonoverridden alert were classified using a data log rather than a medical record review, the alert appropriateness was 7% (14/200) for overridden alerts and 7.7% (14/182) for nonoverridden alerts, which did not show clinical relevance. Contrary to the expectation that there were more inappropriate alerts in nonoverridden alerts, there was no difference in alert appropriateness between the 2 types of alerts (Multimedia Appendix 1).

In the classification of the 382 alerts, only 3.4% (n=13) were successfully triggered, and 2.1% (n=8) were inappropriate for both the alert and physicians' response (Table 4). Only 3.9% (n=15) of alerts represented physicians' nonadherence, where the alert was appropriate but the corresponding physicians' response was inappropriate. The override rate was 92.9% (n=355): *(Physicians' nonadherence [n=15] + justifiable overrides [n=340]) / total alerts [n=382]* (Table 4). There were 6 (1.6%) cases in which the physicians' response could not be determined.

Figure 2. Study flow chart. DDI: drug-drug interaction.

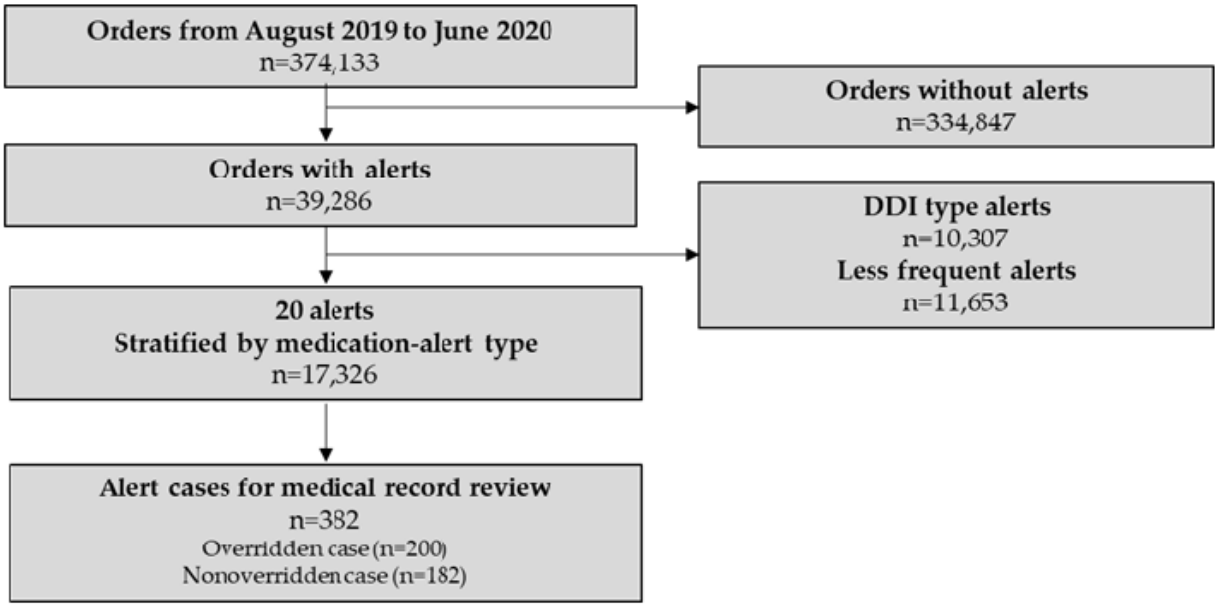


Table 1. The 20 analyzed medication alerts.

Order (medication type)	Alert type	Alert counts, n	Overridden alerts for medical record reviews (N=200), n	Nonoverridden alerts for medical record reviews (N=182), n
Sodium bicarbonate, 8.4%, 20 mL (other)	Dose	2125	10	10
Esomeprazole, 40 mg (proton pump inhibitor)	Dose	1885	10	10
Ceftriaxone sodium, 2 g (antibiotic)	Renal	1379	10	10
Kalimate powder, 5 g (other)	Dose	1494	10	10
Tazoferan, 2.25 g (antibiotic)	Renal	1108	10	10
Calcium gluconate, 2 g/20 mL (calcium)	Dose	1230	10	10
Acetaminophen, 1 g/100 mL (analgesic)	Dose	1527	10	10
Pantoprazole, 40 mg (proton pump inhibitor)	Dose	1059	10	10
Lactulose syrup (other)	Dose	701	10	10
Propacetamol, 1 g (analgesic)	Age	1205	10	10
Methylprednisolone, 4 mg (steroid)	Dose	378	10	6
Ibuprofen, 20 mg/mL (NSAIDs[a])	Dose	611	10	10
Levofloxacin, 750 mg (antibiotic)	Renal	421	10	10
Terlipressin acetate, 1 mg (vasoconstrictor)	Dose	386	10	10
Epinephrine, 1 mg (other)	Dose	340	10	9
Amiodarone, 150 mg (antiarrhythmic)	Dose	329	10	10
Meropenem, 500 mg (antibiotic)	Renal	301	10	10
Ampicillin/sulbactam, 1.5 g (antibiotic)	Dose	271	10	9
Cefazolin, 1 g (antibiotic)	Dose	275	10	6
Cefditoren pivoxil, 100 mg (antibiotic)	Dose	301	10	2

[a]NSAID: nonsteroidal anti-inflammatory drug.

Table 2. Patient demographic.

Demographic	Patient (N=356), n (%)
Sex, n (%)	
Female	152 (42.7)
Male	204 (57.3)
Age (years), n (%)	
0 to 20	58 (16.3)
20 to <40	18 (5.1)
40 to <60	75 (21.1)
≥60	205 (57.6)
KTAS[a] score, n (%)	
1 (most critical)	13 (3.7)
2	51 (14.3)
3	197 (55.3)
4	94 (26.4)
5 (least critical)	1 (0.3)
Injury, n (%)	
Noninjury	68 (19.1)
Injury	288 (80.9)
Disposition, n (%)	
Discharge	121 (34)
Admission	193 (54.2)
General ward (n=193)	165 (85.5)
Intensive care unit (n=193)	28 (14.5)
Transfer	22 (6.2)
Death	20 (5.6)

[a]KTAS: Korean Triage Acuity Scale.

Table 3. Appropriateness of alert clinical relevance and physicians' response.

Appropriateness evaluation	Case (N=382), n (%)		
	Appropriate	Inappropriate	Nondecidable
Alert clinical relevance	28 (7.3)	354 (92.7)	0 (0)
Physicians' response	353 (92.4)	23 (6)	6 (1.6)

Table 4. Evaluation of alerts.

Alert clinical relevance	Physicians' response (N=382), n (%)		
	Appropriate	Inappropriate	Nondecidable
Appropriate	13 (3.4)[a]	15 (3.9)[b, c]	0 (0)
Inappropriate	340 (89)[c]	8 (2.1)[d]	6 (1.6)
Nondecidable	0 (0)	0 (0)	0 (0)

[a]Successful alerts.

[b]Physician's nonadherence.

[c]The override rate (355/382, 92.9%) was determined by the sum of these 2 values divided by the total number of alerts.

[d]Unintended adverse consequences.

Discussion

Principal Findings

In this study, we evaluated the appropriateness of the alerts and physicians' responses to the medication-related passive alert system through a detailed medical record review. We found that only 7.3% of alerts were clinically appropriate, and 6% of alerts resulted in inappropriate responses from physicians. Alert fatigue is inevitable when a large number of irrelevant alerts are generated for a small number of appropriate alerts. There were a few successful alerts where the alert was appropriate and the physician accepted the alert. Physicians' nonadherence of alerts could be a result of the ambiguous contents of alerts that did not provide helpful information [26]. Additionally, a high number of inappropriate alerts could be a reason for physicians' nonadherence [27]. Physicians were less likely to accept alerts as the number of alerts increased, especially for repeated alerts [6]. When considering the cases where the response of the physician was inappropriate, the alerts where the alert was appropriate were almost twice as common as the alerts where the alert was inappropriate. This finding can be explained by habitual override due to numerous inappropriate alerts [28]. A small number of alerts were classified as resulting in unintended adverse consequences. In a few cases, the physicians' response appropriateness could not be determined, because the passive alert system did not collect the override reasons. There were no cases where the appropriateness of the alert could not be determined.

Many studies have identified the appropriateness of override according to the appropriateness of the alert [1,5,15,29,30], but only a few studies have evaluated the response of physicians [31-33]. Duke et al [31] conducted a randomized controlled trial on DDI alert targets to identify medical staff's adherence according to context-enhanced alerting. Strom et al [32] analyzed the unintended effects of a nearly hard-stop CPOE prescribing alert. Understanding the physicians' response to the CDSS is of importance; however, due to the difficulty in analyzing the response, many researchers simply evaluate the appropriateness of the override. Therefore, it is necessary to increase the utility of the CDSS by understanding physicians' responses.

In our previous study, we reported an override rate of 61.9% [22]. However, in this study, we found that the override rate was 92.9%. There are several reasons for this difference. First, in this study, through medical record reviews, it was confirmed that some cases that were previously evaluated as nonoverridden by log data were clinically overridden. The difference between the override rate when simply using log data and the override rate through a medical record review is large, even within the same system. In this study, the patients' overall prescriptions were analyzed through a detailed patient medical record review, and the definition of "override" was expanded. In the previous study, the classification of overridden and nonoverridden alerts was based only on log data [22]. In this study, however, more override was detected by the medical record review than in the previous study. It was confirmed that a substantial number of cases classified as nonoverridden by log data were actually overridden. We found that many physicians prescribed the same prescription that was considered deleted because of an alert. The prescription was considered an override if it was reissued to the same patient within 48 hours of the alert being issued. Therefore, the override rate might be higher in studies that did not identify the nonoverridden alerts [15,29,34,35]. To calculate the override rate properly, it is necessary to establish a mechanism for systematically determining overrides. A standardized definition of override is needed for a detailed analysis and comparison of CDSSs. Furthermore, in this study, we chose the target alerts as alerts that are frequently overridden, so it could be a reason for the high override rate. Additionally, the change of the knowledge base of the CDSS from Medi-Span (Wolters Kluwer Health) to KIMS POC (KIMS Co) may have affected the override rate.

Further research should investigate techniques for improving alert accuracy by using machine learning (ML) and artificial intelligence (AI), analyze the passive CDSS that has not been extensively studied, and explore the causal relationship between the number of alerts and the physicians' responses. Multiple alerts with low clinical relevance reduce physicians' reliance of alerts. Additionally, many unnecessary alerts can lead to alert fatigue and increase the probability of ignoring truly important alerts [2]. It is necessary to improve the clinical relevance of the alert to increase the physician's alert reliance and optimize the alert. ML and AI could be potential solutions. By introducing ML, the rule-based alert system can be improved, and by introducing AI, alerts can be generated according to the individual condition of the patient [36,37]. Despite the promise of technological approaches to drug safety, the risk of mistake will persist if these systems are not carefully applied and heavy attention is not made to building safer systems of care [2]. These considerations are required to reduce needless alerts, improve their clinical relevance, and increase physicians' alert reliance by assessing CDSS consistently.

Limitations

Our study had several limitations. First, it was performed at a single center with ED practices. Second, the evaluation of physicians' response appropriateness may be subjective, because passive alert systems do not collect the override reasons. In addition, we did not confirm the clinical consequences of alerts for unintended adverse consequences. Only the clinical consequences related to the prescription stage were checked, and the dispensing/administration stage was not analyzed.

Conclusions

We evaluated the appropriateness of the alerts and physicians' responses through a detailed medical record review of the medication-related passive alert system. Only by gaining better knowledge of the physicians' overall behavior is it possible to improve the effectiveness of the CDSS. In our study, most alerts did not reflect the clinical situation of each patient; however, the physicians' responses were mostly appropriate. Alert fatigue is unavoidable when a large number of irrelevant alerts are generated in response to a small number of useful alerts. It is necessary to decrease unnecessary alerts, improve their clinical relevance, increase alert reliability, and optimize alerts.

Authors' Contributions

WCC conceived and designed the experiments; HP performed the experiments; MKC, W Jeong, and HC contributed to the experiments; W Jung and JY analyzed the data; and WCC and HP wrote the paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Comparison of alert appropriateness according to overridden and nonoverridden alerts. There was no difference in the appropriateness of clinical relevance between overridden alerts (7% appropriate) and nonoverridden alerts (7.7% appropriate).
[DOCX File , 12 KB - medinform_v10i10e40511_app1.docx]

References

1. Nanji KC, Slight SP, Seger DL, Cho I, Fiskio JM, Redden LM, et al. Overrides of medication-related clinical decision support alerts in outpatients. J Am Med Inform Assoc 2014 May 01;21(3):487-491 [FREE Full text] [doi: 10.1136/amiajnl-2013-001813] [Medline: 24166725]
2. Ranji SR, Rennke S, Wachter RM. Computerised provider order entry combined with clinical decision support systems to improve medication safety: a narrative review. BMJ Qual Saf 2014 Sep 12;23(9):773-780. [doi: 10.1136/bmjqs-2013-002165] [Medline: 24728888]
3. Miller RA, Waitman LR, Chen S, Rosenbloom ST. The anatomy of decision support during inpatient care provider order entry (CPOE): empirical observations from a decade of CPOE experience at Vanderbilt. J Biomed Inform 2005 Dec;38(6):469-485 [FREE Full text] [doi: 10.1016/j.jbi.2005.08.009] [Medline: 16290243]
4. Kuperman GJ, Bobb A, Payne TH, Avery AJ, Gandhi TK, Burns G, et al. Medication-related clinical decision support in computerized provider order entry systems: a review. J Am Med Inform Assoc 2007 Jan 01;14(1):29-40 [FREE Full text] [doi: 10.1197/jamia.M2170] [Medline: 17068355]
5. Poly TN, Islam M, Yang H, Li YCJ. Appropriateness of overridden alerts in computerized physician order entry: systematic review. JMIR Med Inform 2020 Jul 20;8(7):e15653 [FREE Full text] [doi: 10.2196/15653] [Medline: 32706721]
6. Ancker JS, Edwards A, Nosal S, Hauser D, Mauer E, Kaushal R, with the HITEC Investigators. Correction to: effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system. BMC Med Inform Decis Mak 2019 Nov 18;19(1):227 [FREE Full text] [doi: 10.1186/s12911-019-0971-0] [Medline: 31739801]
7. Khairat S, Marc D, Crosby W, Al Sanousi A. Reasons for physicians not adopting clinical decision support systems: critical analysis. JMIR Med Inform 2018 Apr 18;6(2):e24 [FREE Full text] [doi: 10.2196/medinform.8912] [Medline: 29669706]
8. Wright A, McEvoy DS, Aaron S, McCoy AB, Amato MG, Kim H, et al. Structured override reasons for drug-drug interaction alerts in electronic health records. J Am Med Inform Assoc 2019 Oct 01;26(10):934-942 [FREE Full text] [doi: 10.1093/jamia/ocz033] [Medline: 31329891]
9. Hauskrecht M, Batal I, Hong C, Nguyen Q, Cooper GF, Visweswaran S, et al. Outlier-based detection of unusual patient-management actions: an ICU study. J Biomed Inform 2016 Dec;64:211-221 [FREE Full text] [doi: 10.1016/j.jbi.2016.10.002] [Medline: 27720983]
10. Zenziper Straichman Y, Kurnik D, Matok I, Halkin H, Markovits N, Ziv A, et al. Prescriber response to computerized drug alerts for electronic prescriptions among hospitalized patients. Int J Med Inform 2017 Nov;107:70-75. [doi: 10.1016/j.ijmedinf.2017.08.008] [Medline: 29029694]
11. Ariosto D. Factors contributing to CPOE opiate allergy alert overrides. AMIA Annu Symp Proc 2014 Nov 14;2014:256-265 [FREE Full text] [Medline: 25954327]
12. Chaparro JD, Hussain C, Lee JA, Hehmeyer J, Nguyen M, Hoffman J. Reducing interruptive alert burden using quality improvement methodology. Appl Clin Inform 2020 Jan 15;11(1):46-58 [FREE Full text] [doi: 10.1055/s-0039-3402757] [Medline: 31940671]
13. Phansalkar S, van der Sijs H, Tucker AD, Desai AA, Bell DS, Teich JM, et al. Drug-drug interactions that should be non-interruptive in order to reduce alert fatigue in electronic health records. J Am Med Inform Assoc 2013 May 01;20(3):489-493 [FREE Full text] [doi: 10.1136/amiajnl-2012-001089] [Medline: 23011124]
14. Hussain M, Reynolds TL, Zheng K. Medication safety alert fatigue may be reduced via interaction design and clinical role tailoring: a systematic review. J Am Med Inform Assoc 2019 Oct 01;26(10):1141-1149 [FREE Full text] [doi: 10.1093/jamia/ocz095] [Medline: 31206159]
15. Rehr CA, Wong A, Seger DL, Bates DW. Determining inappropriate medication alerts from "inaccurate warning" overrides in the intensive care unit. Appl Clin Inform 2018 Apr 25;9(2):268-274 [FREE Full text] [doi: 10.1055/s-0038-1642608] [Medline: 29695013]

16. Orenstein EW, Kandaswamy S, Muthu N, Chaparro JD, Hagedorn PA, Dziorny AC, et al. Alert burden in pediatric hospitals: a cross-sectional analysis of six academic pediatric health systems using novel metrics. J Am Med Inform Assoc 2021 Nov 25;28(12):2654-2660. [doi: 10.1093/jamia/ocab179] [Medline: 34664664]
17. Kizzier-Carnahan V, Artis KA, Mohan V, Gold JA. Frequency of passive EHR alerts in the ICU: another form of alert fatigue? J Patient Saf 2019 Sep;15(3):246-250 [FREE Full text] [doi: 10.1097/PTS.0000000000000270] [Medline: 27331600]
18. Weingart SN, Toth M, Sands DZ, Aronson MD, Davis RB, Phillips RS. Physicians' decisions to override computerized drug alerts in primary care. Arch Intern Med 2003 Nov 24;163(21):2625-2631. [doi: 10.1001/archinte.163.21.2625] [Medline: 14638563]
19. Chisholm CD, Collison EK, Nelson DR, Cordell WH. Emergency department workplace interruptions: are emergency physicians "interrupt-driven" and "multitasking"? Acad Emerg Med 2000 Nov;7(11):1239-1243 [FREE Full text] [doi: 10.1111/j.1553-2712.2000.tb00469.x] [Medline: 11073472]
20. Forster AJ, Rose NGW, van Walraven C, Stiell I. Adverse events following an emergency department visit. Qual Saf Health Care 2007 Feb 01;16(1):17-22 [FREE Full text] [doi: 10.1136/qshc.2005.017384] [Medline: 17301197]
21. Scheepers-Hoeks AJ, Grouls RJ, Neef C, Ackerman EW, Korsten EH. Physicians' responses to clinical decision support on an intensive care unit--comparison of four different alerting methods. Artif Intell Med 2013 Sep;59(1):33-38. [doi: 10.1016/j.artmed.2013.05.002] [Medline: 23746663]
22. Cha WC, Jung W, Yu J, Yoo J, Choi J. Temporal change in alert override rate with a minimally interruptive clinical decision support on a next-generation electronic medical record. Medicina (Kaunas) 2020 Nov 30;56(12):662 [FREE Full text] [doi: 10.3390/medicina56120662] [Medline: 33265954]
23. McCoy AB, Waitman LR, Lewis JB, Wright JA, Choma DP, Miller RA, et al. A framework for evaluating the appropriateness of clinical decision support alerts and responses. J Am Med Inform Assoc 2012 May 01;19(3):346-352 [FREE Full text] [doi: 10.1136/amiajnl-2011-000185] [Medline: 21849334]
24. Park J, Lim T. Korean Triage and Acuity Scale (KTAS). J Korean Soc Emerg Med 2017 Dec 31;28(6):547-551 [FREE Full text]
25. Kwon H, Kim YJ, Jo YH, Lee JH, Lee JH, Kim J, et al. The Korean Triage and Acuity Scale: associations with admission, disposition, mortality and length of stay in the emergency department. Int J Qual Health Care 2019 Jul 01;31(6):449-455. [doi: 10.1093/intqhc/mzy184] [Medline: 30165654]
26. Shah S, Amato MG, Garlo KG, Seger DL, Bates DW. Renal medication-related clinical decision support (CDS) alerts and overrides in the inpatient setting following implementation of a commercial electronic health record: implications for designing more effective alerts. J Am Med Inform Assoc 2021 Jun 12;28(6):1081-1087 [FREE Full text] [doi: 10.1093/jamia/ocaa222] [Medline: 33517413]
27. Getty DJ, Swets JA, Pickett RM, Gonthier D. System operator response to warnings of danger: a laboratory investigation of the effects of the predictive value of a warning on human response time. J Exp Psychol Appl 1995 Mar;1(1):19-33. [doi: 10.1037/1076-898x.1.1.19]
28. Baysari MT, Tariq A, Day RO, Westbrook JI. Alert override as a habitual behavior - a new perspective on a persistent problem. J Am Med Inform Assoc 2017 Mar 01;24(2):409-412 [FREE Full text] [doi: 10.1093/jamia/ocw072] [Medline: 27274015]
29. Wong A, Seger DL, Slight SP, Amato MG, Beeler PE, Fiskio JM, et al. Evaluation of 'definite' anaphylaxis drug allergy alert overrides in inpatient and outpatient settings. Drug Saf 2018 Mar 9;41(3):297-302. [doi: 10.1007/s40264-017-0615-1] [Medline: 29124665]
30. Stultz JS, Nahata MC. Appropriateness of commercially available and partially customized medication dosing alerts among pediatric patients. J Am Med Inform Assoc 2014 Feb 01;21(e1):e35-e42 [FREE Full text] [doi: 10.1136/amiajnl-2013-001725] [Medline: 23813540]
31. Duke JD, Li X, Dexter P. Adherence to drug-drug interaction alerts in high-risk patients: a trial of context-enhanced alerting. J Am Med Inform Assoc 2013 May 01;20(3):494-498 [FREE Full text] [doi: 10.1136/amiajnl-2012-001073] [Medline: 23161895]
32. Strom BL, Schinnar R, Aberra F, Bilker W, Hennessy S, Leonard CE, et al. Unintended effects of a computerized physician order entry nearly hard-stop alert to prevent a drug interaction: a randomized controlled trial. Arch Intern Med 2010 Sep 27;170(17):1578-1583. [doi: 10.1001/archinternmed.2010.324] [Medline: 20876410]
33. Taegtmeyer AB, Kullak-Ublick GA, Widmer N, Falk V, Jetter A. Clinical usefulness of electronic drug-drug interaction checking in the care of cardiovascular surgery inpatients. Cardiology 2012 Nov 27;123(4):219-222 [FREE Full text] [doi: 10.1159/000343272] [Medline: 23208189]
34. Slight SP, Beeler PE, Seger DL, Amato MG, Her QL, Swerdloff M, et al. A cross-sectional observational study of high override rates of drug allergy alerts in inpatient and outpatient settings, and opportunities for improvement. BMJ Qual Saf 2017 Mar 18;26(3):217-225 [FREE Full text] [doi: 10.1136/bmjqs-2015-004851] [Medline: 26993641]
35. Cho I, Slight SP, Nanji KC, Seger DL, Maniam N, Dykes PC, et al. Understanding physicians' behavior toward alerts about nephrotoxic medications in outpatients: a cross-sectional analysis. BMC Nephrol 2014 Dec 15;15(1):200 [FREE Full text] [doi: 10.1186/1471-2369-15-200] [Medline: 25511564]

36. Poly TN, Islam M, Muhtar MS, Yang H, Nguyen PAA, Li YCJ. Machine learning approach to reduce alert fatigue using a disease medication-related clinical decision support system: model development and validation. JMIR Med Inform 2020 Nov 19;8(11):e19489 [FREE Full text] [doi: 10.2196/19489] [Medline: 33211018]
37. Rozenblum R, Rodriguez-Monguio R, Volk LA, Forsythe KJ, Myers S, McGurrin M, et al. Using a machine learning system to identify and prevent medication prescribing errors: a clinical and cost analysis evaluation. Jt Comm J Qual Patient Saf 2020 Jan;46(1):3-10. [doi: 10.1016/j.jcjq.2019.09.008] [Medline: 31786147]

Abbreviations

AI: artificial intelligence
CDSS: clinical decision support system
CDW: clinical data warehouse
CPOE: computerized physician order entry
DARWIN: Data Analytics and Research Window for Integrated Knowledge
DDI: drug-drug interaction
ED: emergency department
EMR: electronic medical record
KTAS: Korean Triage and Acuity Scale
ML: machine learning

Edited by C Lovis; submitted 26.06.22; peer-reviewed by WY Zheng, DY Kang; comments to author 31.07.22; revised version received 13.09.22; accepted 18.09.22; published 04.10.22.

Please cite as:
Park H, Chae MK, Jeong W, Yu J, Jung W, Chang H, Cha WC
Appropriateness of Alerts and Physicians' Responses With a Medication-Related Clinical Decision Support System: Retrospective Observational Study
JMIR Med Inform 2022;10(10):e40511
URL: https://medinform.jmir.org/2022/10/e40511
doi:10.2196/40511
PMID:36194461

Original Paper

Evaluating the Impact on Clinical Task Efficiency of a Natural Language Processing Algorithm for Searching Medical Documents: Prospective Crossover Study

Eunsoo H Park[1,2*], BMedSci; Hannah I Watson[2*], BMedSci, MBChB, MSc; Felicity V Mehendale[3], MBBS, MS; Alison Q O'Neil[2,4], BSc, MEng, EngD; Clinical Evaluators[5]

[1]Edinburgh Medical School, College of Medicine and Veterinary Medicine, University of Edinburgh, Edinburgh, United Kingdom

[2]Canon Medical Research Europe, Edinburgh, United Kingdom

[3]Centre for Global Health Research, Usher Institute, University of Edinburgh, Edinburgh, United Kingdom

[4]School of Engineering, University of Edinburgh, Edinburgh, United Kingdom

[5]see Acknowledgements, United Kingdom, United Kingdom

[*]these authors contributed equally

Corresponding Author:
Eunsoo H Park, BMedSci
Edinburgh Medical School
College of Medicine and Veterinary Medicine
University of Edinburgh
The Chancellor's Building
49 Little France Crescent
Edinburgh, EH16 4SB
United Kingdom
Phone: 44 1312426792
Email: e.park-7@sms.ed.ac.uk

Abstract

Background: Information retrieval (IR) from the free text within electronic health records (EHRs) is time consuming and complex. We hypothesize that natural language processing (NLP)–enhanced search functionality for EHRs can make clinical workflows more efficient and reduce cognitive load for clinicians.

Objective: This study aimed to evaluate the efficacy of 3 levels of search functionality (no search, string search, and NLP-enhanced search) in supporting IR for clinical users from the free text of EHR documents in a simulated clinical environment.

Methods: A clinical environment was simulated by uploading 3 sets of patient notes into an EHR research software application and presenting these alongside 3 corresponding IR tasks. Tasks contained a mixture of multiple-choice and free-text questions. A prospective crossover study design was used, for which 3 groups of evaluators were recruited, which comprised doctors (n=19) and medical students (n=16). Evaluators performed the 3 tasks using each of the search functionalities in an order in accordance with their randomly assigned group. The speed and accuracy of task completion were measured and analyzed, and user perceptions of NLP-enhanced search were reviewed in a feedback survey.

Results: NLP-enhanced search facilitated more accurate task completion than both string search (5.14%; *P*=.02) and no search (5.13%; *P*=.08). NLP-enhanced search and string search facilitated similar task speeds, both showing an increase in speed compared to the no search function, by 11.5% (*P*=.008) and 16.0% (*P*=.007) respectively. Overall, 93% of evaluators agreed that NLP-enhanced search would make clinical workflows more efficient than string search, with qualitative feedback reporting that NLP-enhanced search reduced cognitive load.

Conclusions: To the best of our knowledge, this study is the largest evaluation to date of different search functionalities for supporting target clinical users in realistic clinical workflows, with a 3-way prospective crossover study design. NLP-enhanced search improved both accuracy and speed of clinical EHR IR tasks compared to browsing clinical notes without search. NLP-enhanced search improved accuracy and reduced the number of searches required for clinical EHR IR tasks compared to direct search term matching.

(JMIR Med Inform 2022;10(10):e39616) doi:10.2196/39616

KEYWORDS

clinical decision support; electronic health records; natural language processing; semantic search; clinical informatics

Introduction

Background

The benefits of the transition from storing patient information in paper notes to electronic health records (EHRs) have been a topic of debate among health care professionals [1-4]. Many clinicians have expressed dissatisfaction with their current hospital systems, and EHR use is consistently cited as a contributor to clinician burnout [5-7]. Approximately 40% of doctors' time is spent documenting patient information, with evidence showing that this work burden has increased following EHR implementation [8,9]. However, difficulties in quickly and accurately retrieving relevant information from these documents indicate that this wealth of collected information is often not fully used [10,11]. Navigating EHR documents is challenging owing to the complexity of medical text, which tends to include frequent misspellings, abbreviations, specialty-specific acronyms, and clinical shorthand [12-15]. Time-consuming and inaccurate information gathering from EHRs limits the efficiency of wider clinical workflows [16], with some doctors believing that difficulties in retrieving patient information significantly impact face-to-face patient care [17].

Despite the increasing sophistication of general search engines, there remain relatively limited search options within medical record software. One barrier is the need for patient data to be held securely; therefore, access to computing power and shared resources may be limited. To have clinical utility, search facilities must be fast and intuitive for use by time-pressured clinicians, including relatively junior members of staff to whom the task of searching through complex notes is frequently delegated. In addition, the search must handle high variability of text expression as mentioned above. Clinical text is error prone; unlike journals and other publications, there is no editorial control to check for errors. Medical terminology, acronyms, and abbreviations vary between regions and hospitals and even across different specialties; for instance, "CHD" may be related to chronic heart disease (cardiology), congenital heart disease (pediatrics), or congenital hip dislocation (orthopedics). Since clinical care is a high-stakes environment, failure to find relevant information potentially has great implications; to effectively save the time of clinicians, search tools should ideally go beyond document-level results to locate and highlight all relevant sentences or even words within a document. Efforts to achieve easier information retrieval (IR) have included the integration of string search in some EHRs, similar to the "Ctrl-F" or "Find" function that is now frequently available on everyday platforms [18]. However, the effectiveness of string search is limited for heterogeneous clinical text; therefore, studies have also considered semantic search algorithms [19-22]. A large-scale retrospective analysis of searches performed in an EHR found that the use of search varied considerably across and within user roles, with physicians and pharmacists being the most active user groups [19]. A review of the use of search within EHRs found that few articles focused on the impact of search within clinical workflows [23]; one study with 7 diabetes experts found that content-based search was both faster and more accurate than conventional search for finding relevant information [20], another study with 10 family and internal medicine physicians found that semantic search allowed for faster medical notes navigation for IR tasks [21], and a final study with 4 students found that a semantic search tool enabled faster clinical note summarization [22]. Only one of the described studies [20] used a crossover study design. In this paper, a larger study is reported (n=35 valid task completions, n=42 qualitative responses), in which a 3-way prospective crossover study was conducted, comparing a standard string search with no search and with a natural language processing (NLP)–enhanced search. The custom NLP-enhanced search tool combines ontologies with fuzzy matching to offer search functionality, which captures not only semantically related terms (eg, synonyms and hyponyms) but also linguistic alternative spellings and misspellings and word forms of the search term. A simulated clinical environment was used alongside target user feedback to determine whether search tools could make clinical workflows more efficient and reduce clinicians' cognitive burdens when attempting to find information.

Aims and Hypotheses

This study aimed to quantitatively and qualitatively compare the efficacy of 3 search functionalities for IR from medical free-text documents, in terms of their accuracy, speed, and ease of search.

We hypothesized that search tools will allow clinical users to perform simulated clinical IR tasks faster and more accurately than when using no search, with the use of NLP techniques enabling NLP-enhanced search to perform more effectively than string search.

Methods

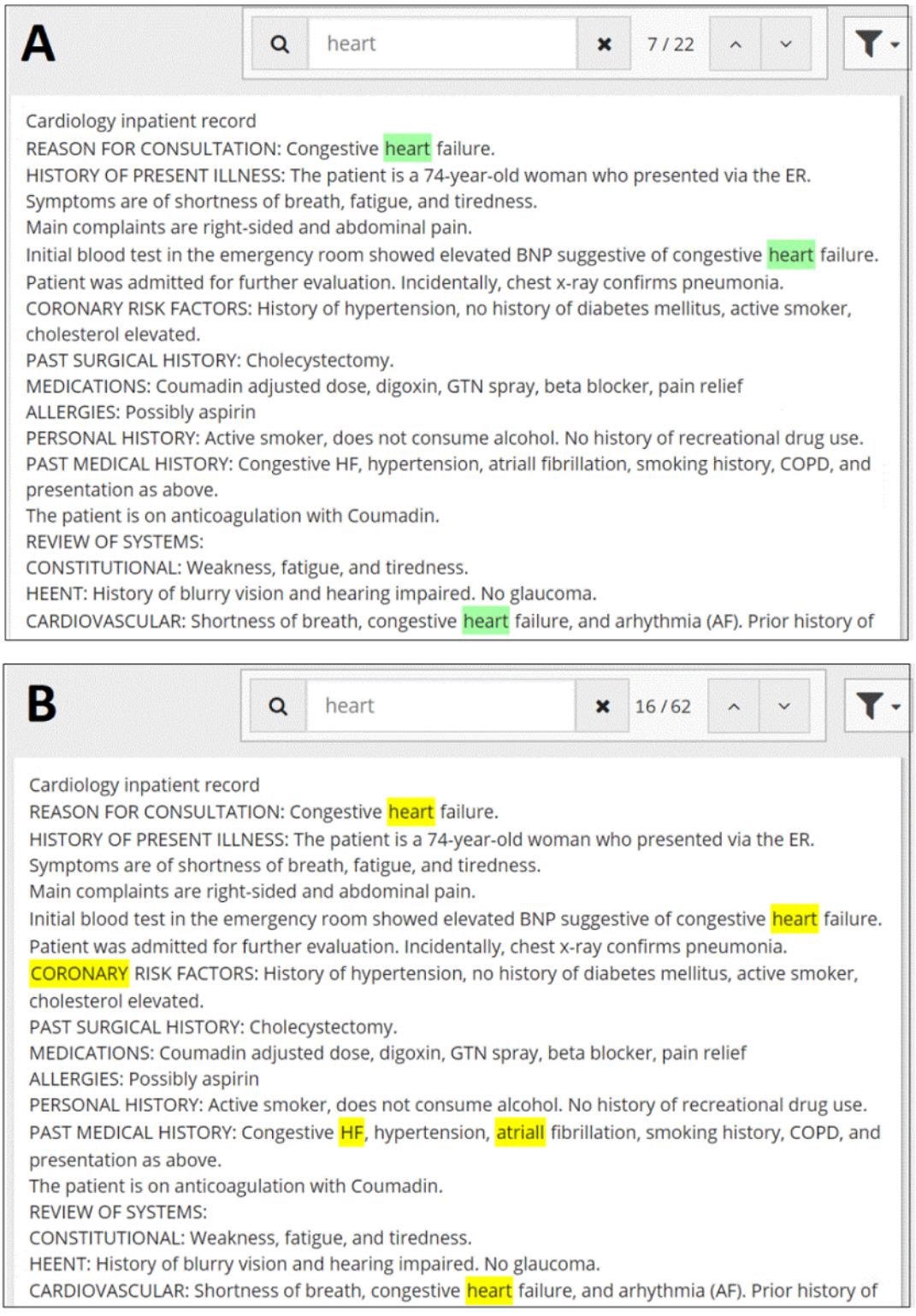
Search Tools

The string search function is an open source JavaScript library implementation [24]. NLP-enhanced search is a proprietary rule-based algorithm (developed at Canon Medical Research Europe) that leverages NLP techniques such as edit distance and stemming in conjunction with medical knowledge bases, notably the Unified Medical Language System semantic network, Metathesaurus [25], and medical abbreviation lists on Wikipedia [26] and OpenMD [27]. These sources are used to expand the original search term into a list of equivalent terms, which are then located in the text. The tool was designed to locate linguistic variants such as misspellings and alternative spellings, word forms, and abbreviations, as well as additional semantic synonyms.

Search tools were integrated into a patient-centric viewer (EHR research software), which allowed the user to type in a search

term and view the highlighted findings within the retrieved subset of documents, which the user could scroll through. In the case of no search, the user was expected to scroll through the patient's EHR to find the relevant information. Figure 1 illustrates the difference between the two search tools in the patient-centric viewer.

Figure 1. Example results for (A) string search and (B) NLP-enhanced search for the search term "heart." String search returned only direct matches to "heart" (green highlights) whereas NLP-enhanced search also returns semantically related terms (yellow highlights) such as the following: "coronary," the misspelling of atrial (fibrillation) as "atriall," and the appearance of "heart" within the abbreviation of heart failure, "HF." NLP: natural language processing.



Simulating a Clinical Workflow

Overview

Free-text medical documents were synthesized for 3 fictional patients. These materials were paired with corresponding sets of 10 IR questions for each patient, grounded in relevant and realistic clinical scenarios. Patient documents were uploaded into the patient-centric viewer. Questions were uploaded onto a custom evaluation platform built using Vue.js, which also displayed the clinical scenarios and task-specific instructions for the evaluator. Below, we describe the document synthesis and question generation in more detail.

Patient Document Synthesis

Three patient profiles were created with varying age, sex, ethnic background, social history, and medical history. The 3 patients were assigned primary medical specialties of respiratory, neurology, and oncology. For each patient, 20 documents were created by selecting and augmenting publicly available anonymized medical documents [28], as well as manually synthesizing additional documents to provide a patient EHR with a coherent chronological sequence of clinical events. Documents were varied and included discharge letters, outpatient clinic letters, operation notes, and general practice referral letters. To imitate real-world medical text, common misspellings, abbreviations, and acronyms were included in the text, using investigator clinical experience (author HW) and reference papers [13].

Clinical Scenarios and Question Generation

For each task, clinical scenarios were designed to focus on real-world situations where information can be extracted from patient notes. To ensure that the tasks were comparable across patients (and therefore interventions), a master template of 10 questions prompting IR was created, which was then tailored to fit each patient scenario. Questions were inspired by those in past medical examinations [29] and investigators' (HW and FM) clinical experience. Requested information resembled that required in typical clinical workflows to support clinical decision-making. Care was taken to ensure that task questions tested the search function and not clinical knowledge or judgement; therefore, all answers could be found by searching the respective patient's notes. Questions required a mixture of multiple-choice and free-text responses. Examples of scenarios and corresponding questions for each patient can be seen in Table 1.

Table 1. Examples of clinical scenarios for each patient and their corresponding question-answer options. Scenarios aimed to simulate a standard clinical workflow, providing context for the questions.

Patient	Example clinical scenario	Example question	Answer type

1	You're worried this may be an exacerbation of a previously present infection. After sending the patient for a chest X-ray and taking bloods, you continue to search for more information.	Does this patient have a history of respiratory infection during the months December 2020-February 2021?	• Select one of the following:
			  • Yes
			  • No
			  • Information not available

		Why was the patient's nitrofurantoin stopped?	• Free text

2	Patient presents to the Emergency Department with confusion and acute stroke-like symptoms. His son reports 2 previous "mini-strokes". You are an ED registrar and send him for a CT head, as per protocol. While waiting for the results you search his history for other contraindications to thrombolysis treatment.	Does the patient have a history of head trauma or stroke between November 2020 and February 2021 (inclusive)?	• Select one of the following:
			  • Yes
			  • No
			  • Information not available

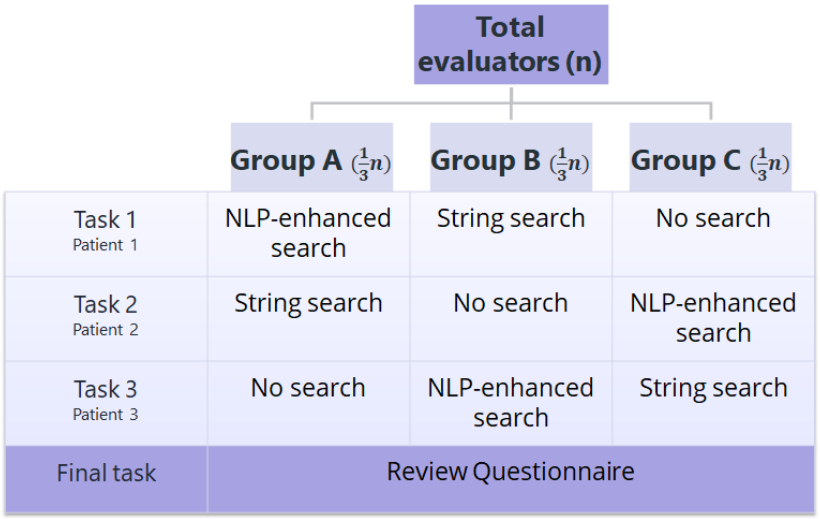
		Search the notes to find the dates of the aforementioned "mini-strokes" (e.g. transient ischaemic attacks).	• Free text
3	You are the new oncologist at the clinic seeing this patient for review. Prior to the appointment you want to check her history by accessing her notes so you can adequately prepare yourself for the consultation.	What is the patient's cancer diagnosis?	• Free text

		Does this patient have a history of any of the following conditions?	• Select all that apply:
			  • Metastases
			  • Hypertension
			  • Epilepsy
			  • Asthma
			  • None of the above

Study Design

The clinical evaluation pipeline was structured as having a prospective crossover trial design; we have illustrated this in Figure 2. Evaluators were banded on the basis of their level of clinical experience before being assigned pseudonymized evaluator IDs that were used for the remainder of the study and analysis. Evaluators in each band were then randomly allocated across the 3 study groups using a random number generator. This yielded 3 groups stratified for level of clinical experience. Each group had a predetermined order of search functionality; once the 3 tasks were completed using the allocated search order, evaluators were asked to fill out a feedback survey that focused on their user experience.

Figure 2. Study design. The 3 tasks were performed using a prospective crossover design, in which each group undertook the tasks in the same order with a predetermined order of the search intervention; the order was different for different groups. Finally, all evaluators were asked to fill in a review questionnaire. NLP: natural language processing.



Evaluator Recruitment and Training

Recruitment for the study was accomplished via professional contacts and advertising on social media channels to reach evaluators from a variety of clinical specialties and years of clinical experience.

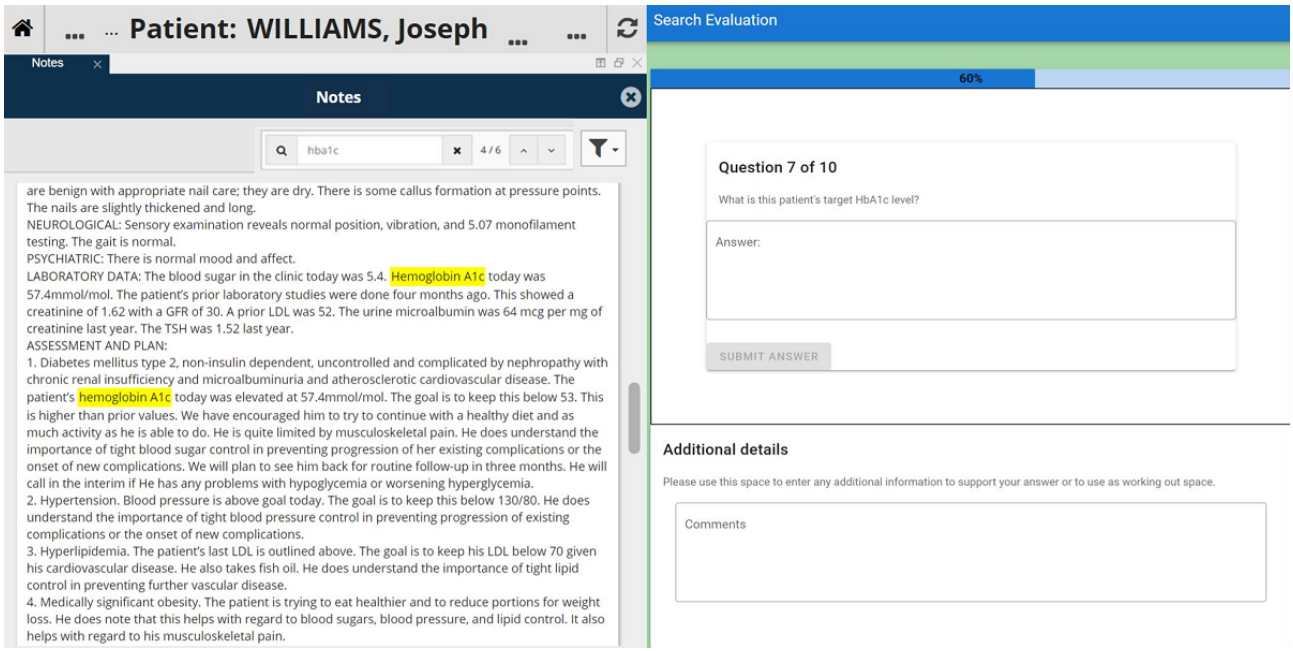
A training video was provided to evaluators, which comprised a brief introduction to the study, demonstrations of the 3 search interventions within the patient-centric viewer, and detailed instructions on how to complete the evaluation. An example patient with a small set of curated medical documents was also provided for training, on which evaluators could familiarize themselves with the capabilities of the different search functionalities.

Data Collection

Evaluators were provided with secure remote access to the evaluation environment (Figure 3), allowing the evaluation to be performed remotely from personal devices. Using this setup, evaluators could view the patient-centric viewer and the evaluation platform. Answers had to be inputted sequentially on the evaluation platform, which did not allow evaluators to return to a question once they had submitted an answer.

During each task, the evaluators submitted answers to the task questions through the evaluation platform. To ensure accurate recording of task times, evaluators were asked to perform each task in one go and to take breaks between tasks rather than during tasks. Evaluators were free to spend as long as they needed on each task. In addition, a search log was maintained, which recorded the search terms entered by the evaluator along with the search functionality used along with the time spent on each question.

Figure 3. Screenshot of the evaluation environment during a task. Evaluators only had permission to view the two relevant sites: the patient-centric viewer (left) and the evaluation platform (right). The patient-centric viewer contains the synthetic patient documents for a given patient (in this case "Joseph Williams") with "hba1c" as the search term. The evaluation platform detailed the clinical scenarios, task-specific instructions, and question-and-answer sections.



Data Analysis

Exclusion Criteria

Data were excluded where search logs showed that evaluators had used an incorrect search functionality for a given task.

Question Marking

Two clinical investigators (EP and HW) reached a consensus on the correct answers for each question. Answers were then clustered depending on the document in which they were located, and marks were awarded for finding each relevant area of correct documents. For example, if 3 pieces of clinical information across 2 unique documents were required to correctly answer a question, then 2 marks were awarded if the correct answer was inputted as the evaluator had successfully found both documents. Questions were weighted equally.

Statistical Analysis

Data analysis was performed using custom Python code. For all metrics, samples were weighted to compensate for imbalances in group size (see *Evaluator Demographics and Group Stratification*). Paired 2-tailed *t* tests were performed to determine if there was a significant difference in timing and accuracy between (1) string search and no search, (2) NLP-enhanced search and no search, and (3) NLP-enhanced search and string search. A significance level of P=.10 was applied.

Search Term Analysis

Following the study, search term logs were analyzed to extract the number and pattern of search terms for each type of search.

User Perceptions

User perceptions were assessed via a feedback survey (see Multimedia Appendix 1) which included a mix of Likert scale ratings, from "strongly disagree" to "strongly agree," and free-text responses. We clustered free-text responses by topic; we have summarized our overall findings in the *User Perceptions of NLP-Enhanced Search* section as they relate to 4 underlying questions of interest: "How was NLP-enhanced search perceived?"; "Is NLP-enhanced search better than string search?"; "Would NLP-enhanced search make clinical workflows more efficient?"; and "Would NLP-enhanced search reduce cognitive load?"

Results

Evaluator Demographics and Group Stratification

In total, 60 evaluators were recruited with multiple levels of clinical experience from medical students to doctors and from 9 specialties ranging from vascular surgery to general practice. Of 60 recruited evaluators, 44 completed the tasks; 35 were included in the final analysis (Table 2), while 9 were excluded. Evaluators were excluded from the quantitative analysis if their data were corrupted (n=2) or they completed the tasks incorrectly (n=7); for example, by using the wrong search functionality for a given task. From the original 20 evaluators per group, we observed 7 (group 1), 13 (group 2), and 15 (group 3) successful completions. There were 42 responses to the feedback survey. Table 2 shows the final distribution of clinical experience across the groups.

Table 2. Summary of allocation across clinical bands and study groups.

Clinical band	Group 1	Group 2	Group 3	Total
Medical students, n				
Preclinical (years 1-3)	4	3	3	10
Clinical (years 4-6)	0	4	2	6
Doctors, n				
1-5 years of clinical experience	0	3	3	6
6-10 years of clinical experience	1	1	4	6
11+ years of clinical experience	2	2	3	7
Total, n	7	13	15	35

Effect of Search Functionality on the Speed and Accuracy of Task Completion

The results are shown in Tables 3 and 4. Overall, NLP-enhanced search facilitated significantly more accurate task completion than both string search (5.14%) and no search (5.13%). In terms of speed, NLP-enhanced search and string search facilitated significantly faster task completion than no search (11.5% and 16.0%, respectively); there was no significant time difference between string search and NLP-enhanced search.

Table 3. Accuracy and time for different search functionalities, showing mean (SD) values across evaluators.

Search functionality	Accuracy (%), mean (SD)	Time per task (minutes), mean (SD)
None	83.8 (9.94)	20.2 (10.8)
String	83.7 (10.8)	17.0 (5.9)[a]
Natural language processing–enhanced	88.1 (9.07)[a]	17.9 (7.20)

[a]Best outcomes.

Table 4. Pairwise comparisons among different search functionalities, showing mean (SD) values in the difference across evaluators.

Search functionality comparison pairs	Accuracy increase (%)		Time difference (minutes)	
	Difference, mean (%; SD)	*P* value	Difference, mean (%; SD)	*P* value
None vs string	–0.01 (0.01; 14.5)	.93	–3.22 (–16.0; 9.78)	.006
None vs NLP-enhanced	4.30 (5.13; 13.1)	.08	–2.32 (–11.5; 7.64)	.008
String vs NLP-enhanced	4.30 (5.14; 10.5)	.02	0.91 (5.34; 5.05)	.18

Analysis of Search Terms Used by Evaluators

Analysis of the logged search terms (Table 5) revealed that evaluators tried almost twice as many search terms when using string search compared to NLP-enhanced search, and uptake of string search was slightly lower than that of NLP-enhanced search; that is, the percentage of questions for which no searches were performed was higher for string search.

The higher number of search terms required for string search might intuitively be explained by the user needing to attempt multiple synonyms to find relevant information. For instance, for the question, “Does the patient have a history of stroke?” in the text, there were 4 negative mentions scattered through the documents: “does not look like she has a stroke,” “No TIA or CVA” (ie, no transient ischemic attack or cerebrovascular accident), “No CVA,” and “No CVA.” NLP-enhanced search found all mentions with the search term “stroke” (which was the only term that evaluators attempted), but string search evaluators also attempted “TIA,” “CVA,” “neurological,” “history,” and “infarction” in their efforts to find all relevant information. Interestingly, we see that evaluators were sometimes searching for neighboring words (“history” or “neurological”) most likely as a method to bypass the possible variation in textual mentions. Further, string search does not match spelling variants (or misspellings); therefore, evaluators sometimes tried different spellings; for example, for the question, “Is the patient currently on full-dose anticoagulant treatment?” both “anti-coagulant” and “anticoagulant” were used as successive search terms by evaluators using string search.

This analysis also highlighted that the strict parameter settings for string search meant that search terms matched only to whole words, not to substrings; thus, evaluators could not search with a prefix. We observed some evidence of evaluators adjusting to this—for example, searching first for “anticoag” and then “anticoagulant” or searching for both “smoke” and “smoker”—and this also increases the number of search terms attempted.

Table 5. Analysis of used search terms showing the percentage of answers that used search and the mean (SD) values of the number of search terms for each of these answers.

Search functionality	Answers using the search functionality, %	Search terms per answer, mean (SD)
String	83.7	3.51 (2.91)
Natural language processing–enhanced	95.1[a]	2.05 (1.49)[a]

[a]Best outcomes.

User Perceptions of NLP-Enhanced Search

We used the survey shown in Multimedia Appendix 1 to gather information about user perceptions of NLP-enhanced search. Below we summarize responses under 4 headings.

How Was NLP-Enhanced Search Perceived?

Most respondents positively described the capabilities of NLP-enhanced search, noting its identification of misspellings, word forms, and synonyms, though some reported that NLP-enhanced search returned too many findings ("[NLP-enhanced] search was very clever and thorough but could return 100 results"). However, when rating the efficacy of NLP-enhanced search, 76% of respondents thought that any unrelated findings—that is, false positives—did not significantly impact the usefulness of the search algorithm.

Is NLP-Enhanced Search Better Than String Search?

Overall, 81% of respondents agreed that NLP-enhanced search facilitated more relevant IR than string search. However, many commented that the string search capabilities within the patient-centric viewer were more limited than they were accustomed to on everyday devices, stating that "string search was too discriminatory" (the parameter settings meant that only whole word matches were returned, not substrings, as discussed in the *Analysis of Search Terms Used by Evaluators* section).

Would NLP-Enhanced Search Make Clinical Workflows More Efficient?

Overall, 93% (39/42) respondents agreed that NLP-enhanced search would make clinical workflows more efficient than string search, in particular during clinics and clerking of patients. Free-text feedback reflected this, with respondents reporting that NLP-enhanced search was useful and less time consuming than string search or no search when retrieving specific information. One evaluator commented, "the [NLP-enhanced] search tool made it significantly easier for me to find the information I was looking for and also quicker." On the other hand, respondents further reported that NLP-enhanced search would not always be the best method for situations where a comprehensive overview of a patient is needed. In this case, assimilating information using manual review (no search) would be more effective. One evaluator said, "I felt that using the [NLP-enhanced] search tool meant I wasn't focussing on the case as much but just looking for words." A common opinion was that NLP-enhanced search would be a useful addition to manual review for clinical tasks.

Would NLP-Enhanced Search Reduce the Cognitive Load?

Respondents frequently mentioned that NLP-enhanced search made it easier to retrieve the information they were looking for, with one evaluator stating that "[NLP-enhanced] search is an excellent tool for a quick way to filter through relevant information." While a few mentioned that too many results were returned, respondents also reported that going through the relevant findings was easier and preferable to a full manual review of the notes, with manual review being described as "tedious," "painstaking," and "very easy to miss vital information." One evaluator commented that NLP-enhanced search could "improve the workload of an already overworked profession."

Discussion

Principal Findings

Our results showed a significant increase in accuracy when NLP-enhanced search was used compared to when string search and no search were used, while both NLP-enhanced search and string search offered time savings. There was a perception of easier navigation from evaluators and a measured decrease in required interactivity in the case of NLP-enhanced search (lower number of search terms than those obtained with string search). We caveat this conclusion with the observation that the strict parameter settings of string search meant that search terms matched only with whole words, not substrings; this increased the number of terms that evaluators used and potentially reduced the search accuracy, compared to a string search version that matches also to substrings.

There is limited literature on the potential impact of EHR search tools on day-to-day clinical care [30]. Our results support those of previous studies [20-22], which have reported that semantic search tools allow faster and more accurate EHR task completion in simulated clinical workflows. A related study [31] reported that artificial intelligence–optimized patient records improve speed in answering clinical questions while maintaining the same accuracy. Interestingly, the impact of the patient record search engine MorphoSaurus has been measured in a real-world clinical setting [32], albeit with user surveys only. This method would have had the benefit of involving real-world stresses such as task interruptions and time pressure, as well as the key element of patient interaction. Importantly, however, our method of using a controlled simulated clinical environment enabled us to control for variables such as distractions or interruptions, as well as variation in the complexity and length of medical records. Additionally, our crossover design controlled for individual participants' ability, experience, and diligence. This enabled robust comparison of quantitative and qualitative data

for each search type while minimizing the impact of confounding factors.

Overall, evaluator feedback suggested that the optimum approach to navigating clinical notes is a hybrid of manual browsing and search, depending on the context. In the real world, NLP-enhanced search is likely best employed as a complementary tool to aid clinical users in navigating clinical notes, with the ability to manually parse and ingest relevant facts from a complex medical history remaining important.

Conclusions

In conclusion, this study suggests that search tools have a positive effect on both the measured and perceived accuracy and ease of clinical IR. Search tools that can leverage NLP techniques are more effective for retrieving all relevant terms from heterogeneous medical free text. There is potential to reduce clinicians' cognitive burden and make clinical workflows more efficient. A critical direction for future research is to assess the use of search tools in real-world clinical practice.

Acknowledgments

We thank Prof Keith W Muir (Institute of Neuroscience & Psychology, University of Glasgow) for his clinical insights during the development of the natural language processing (NLP)-enhanced search tool. We would like to thank the West of Scotland Safe Haven within National Health Service (NHS) Greater Glasgow and Clyde for assistance in creating and providing a data set that was used during development of the NLP-enhanced search tool.

Many thanks to the Canon Medical Research Europe staff who developed the infrastructure required for this evaluation: Yvonne Belton, Michael Corrigan, Vismantas Dilys, Francisco Gomez, Graham Jones, Hamish MacKinnon, David Miller, Emel Muzaç, Paul Norman, and Euan Robertson. Further, we would like to acknowledge the research team that was responsible for creating the NLP-enhanced search tool: Murray Cutforth, Vismantas Dilys, Matúš Falis, Aneta Lisowska, Hamish MacKinnon, Maciej Pajak, Alison O'Neil, and Hannah Watson.

We thank our evaluators: Fiona Auld, Anna Barton, Rong Bing, Cameron Brown, Khai Syuen Chew, Jane Yi Chiam, Vanessa Chou, Luisa Ciriello, George Cooper, Iona Cutworth, Jamie Donachie, Vivienne Evans, Magdalena Gabrysiak, Eilidh Gunn, Mohamed Hamed, Hamzah Hanif, Ewen Harrison, Kylla Hernandez, Lana Huang, Katie Hunter, Haider Khan, David Kluth, Niki Kouvroukoglou, Barbora Krivankova, Tommy Le, Charles Leeson-Payne, Alinah Sum-Ping Leung, Jenny Lockhart, Jack Lugue, Angus MacLeod, Tomos Morgan, Ellen Murgitroyd, Sarah Murphy, Helen O'Neil, Yusuke Onishi, Lisa Ragunathan, Nikita Rana, Qi Shun Yong, Lucy Taylor, Evangelos Tzolos, Miriam Veenhuizen, Philippa Veenhuizen, Olivia Yu, and Sydney Zides.

We thank our pretrial evaluators: Marcus Boyd, Elizabeth Daly, Greta Economides, Keziah Lewis, Abhishek Nambiar, Sumrah Naqvi, Risako Sakatsume, Faye Sikora, and Emma Warburton.

We thank our internal Canon reviewers: William Clackett, Russell Hung, and David Miller.

We thank MTSamples for permitting free use and modification of their data to create the patient case studies.

This work is part of the Industrial Centre for Artificial intelligence (AI) Research in digital Diagnostics, which is funded by Innovate UK on behalf of UK Research and Innovation (project 104690). FV Mehendale's research at the University of Edinburgh is supported by the Caledonian Heritable Foundation.

Authors' Contributions

EHP co-designed the study, co-designed the patient histories, reviewed the synthetic patient notes, designed the tasks, designed the clinical feedback survey, organized evaluator recruitment, recorded training materials for evaluators, performed preliminary analysis of the findings, and contributed to the manuscript draft. HIW co-designed the study, co-designed the patient histories, created the synthetic patient notes, reviewed the tasks, reviewed the clinical feedback survey, supported evaluator recruitment, organized the infrastructure for the practical evaluation, contributed to and reviewed the analysis, and contributed to and reviewed the paper draft. FVM co-designed the study, reviewed the patient histories, reviewed the synthetic patient notes, reviewed the tasks, reviewed the clinical feedback survey, reviewed the analysis, and contributed to and reviewed the paper draft. AQO co-designed the study, organized provision of the NLP-enhanced search, reviewed the tasks, reviewed the clinical feedback survey, contributed to and reviewed the analysis, and contributed to and reviewed the manuscript draft.

Conflicts of Interest

HIW and AQO are employees of Canon Medical Research Europe, who provided the software and algorithms for this evaluation. EHP was sponsored by Canon Medical Research Europe during her Spring 2021 BSc research project at the University of Edinburgh ("Evaluation of a natural language processing algorithm for searching medical documents") which was the basis for this evaluation. EHP had previously performed paid annotation work for the development of the NLP-enhanced search tool.

Multimedia Appendix 1
Feedback survey which the evaluators were requested to fill out on completion of the clinical tasks.

[PDF File (Adobe PDF File), 198 KB - medinform_v10i10e39616_app1.pdf]

References

1. Holanda A, do Carmo E Sá HL, Vieira A, Catrib AMF. Use and satisfaction with electronic health record by primary care physicians in a health district in Brazil. J Med Syst 2012 Oct;36(5):3141-3149 [FREE Full text] [doi: 10.1007/s10916-011-9801-3] [Medline: 22072279]
2. King J, Patel V, Jamoom E, Furukawa MF. Clinical benefits of electronic health record use: national findings. Health Serv Res 2014 Feb;49(1 Pt 2):392-404 [FREE Full text] [doi: 10.1111/1475-6773.12135] [Medline: 24359580]
3. Burke H, Sessums L, Hoang A, Becher DA, Fontelo P, Liu F, et al. Electronic health records improve clinical note quality. J Am Med Inform Assoc 2015 Jan 1;22(1):199-205 [FREE Full text] [doi: 10.1136/amiajnl-2014-002726] [Medline: 25342178]
4. Entzeridou E, Markopoulou E, Mollaki V. Public and physician's expectations and ethical concerns about electronic health record: benefits outweigh risks except for information security. Int J Med Inform 2018 Feb;110:98-107 [FREE Full text] [doi: 10.1016/j.ijmedinf.2017.12.004] [Medline: 29331259]
5. Kroth P, Morioka-Douglas N, Veres S, Babbott S, Poplau S, Qeadan F, et al. Association of electronic health record design and use factors with clinician stress and burnout. JAMA Netw Open 2019 Aug 02;2(8):e199609 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.9609] [Medline: 31418810]
6. Starren J, Tierney W, Williams M, Tang P, Weir C, Koppel R, et al. A retrospective look at the predictions and recommendations from the 2009 AMIA policy meeting: did we see EHR-related clinician burnout coming? J Am Med Inform Assoc 2021 Apr 23;28(5):948-954 [FREE Full text] [doi: 10.1093/jamia/ocaa320] [Medline: 33585936]
7. Yan Q, Jiang Z, Harbin Z, Tolbert PH, Davies MG. Exploring the relationship between electronic health records and provider burnout: a systematic review. J Am Med Inform Assoc 2021 Apr 23;28(5):1009-1021 [FREE Full text] [doi: 10.1093/jamia/ocab009] [Medline: 33659988]
8. Sinsky C, Colligan L, Li L, Prgomet M, Reynolds S, Goeders L, et al. Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties. Ann Intern Med 2016 Sep 06;165(11):753 [FREE Full text] [doi: 10.7326/m16-0961]
9. Joukes E, Abu-Hanna A, Cornet R, de Keizer NF. Time spent on dedicated patient care and documentation tasks before and after the introduction of a structured and standardized electronic health record. Appl Clin Inform 2018 Jan;9(1):46-53 [FREE Full text] [doi: 10.1055/s-0037-1615747] [Medline: 29342479]
10. Beasley J, Wetterneck T, Temte J, Lapin JA, Smith P, Rivera-Rodriguez AJ, et al. Information chaos in primary care: implications for physician performance and patient safety. J Am Board Fam Med 2011;24(6):745-751 [FREE Full text] [doi: 10.3122/jabfm.2011.06.100255] [Medline: 22086819]
11. Blijleven V, Koelemeijer K, Jaspers M. Identifying and eliminating inefficiencies in information system usage: a lean perspective. Int J Med Inform 2017 Nov;107:40-47 [FREE Full text] [doi: 10.1016/j.ijmedinf.2017.08.005] [Medline: 29029690]
12. Meystre S, Savova G, Kipper-Schuler K, Hurdle J. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform 2018 Mar 07;17(01):128-144. [doi: 10.1055/s-0038-1638592]
13. Sinha S, McDermott F, Srinivas G, Houghton PWJ. Use of abbreviations by healthcare professionals: what is the way forward? Postgrad Med J 2011 Jul;87(1029):450-452 [FREE Full text] [doi: 10.1136/pgmj.2010.097394] [Medline: 21459778]
14. Turchin A, Chu JT, Shubina M, Einbinder JS. Identification of misspelled words without a comprehensive dictionary using prevalence analysis. AMIA Annu Symp Proc 2007 Oct 11:751-755 [FREE Full text] [Medline: 18693937]
15. Zhou L, Mahoney LM, Shakurova A, Goss F, Chang FY, Bates DW, et al. How many medication orders are entered through free-text in EHRs?--a study on hypoglycemic agents. AMIA Annu Symp Proc 2012;2012:1079-1088 [FREE Full text] [Medline: 23304384]
16. Farri O, Pieckiewicz DS, Rahman AS, Adam TJ, Pakhomov SV, Melton GB. A qualitative analysis of EHR clinical document synthesis by clinicians. AMIA Annu Symp Proc 2012;2012:1211-1220 [FREE Full text] [Medline: 23304398]
17. Grabenbauer L, Skinner A, Windle J. Electronic health record adoption – maybe it's not about the money. Appl Clin Inform 2017 Dec 16;02(04):460-471 [FREE Full text] [doi: 10.4338/aci-2011-05-ra-0033]
18. Yang L, Mei Q, Zheng K, Hanauer DA. Query log analysis of an electronic health record search engine. AMIA Annu Symp Proc 2011;2011:915-924 [FREE Full text] [Medline: 22195150]
19. Ruppel H, Bhardwaj A, Manickam RN, Adler-Milstein J, Flagg M, Ballesca M, et al. Assessment of electronic health record search patterns and practices by practitioners in a large integrated health care system. JAMA Netw Open 2020 Mar 02;3(3):e200512 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.0512] [Medline: 32142128]
20. Duftschmid G, Rinner C, Kohler M, Huebner-Bloder G, Saboor S, Ammenwerth E. The EHR-ARCHE project: satisfying clinical information needs in a shared electronic health record system based on IHE XDS and archetypes. Int J Med Inform 2013 Dec;82(12):1195-1207 [FREE Full text] [doi: 10.1016/j.ijmedinf.2013.08.002] [Medline: 23999002]
21. Tawfik A, Kochendorfer K, Saparova D, Al Ghenaimi S, Moore JL. Using semantic search to reduce cognitive load in an electronic health record. 2011 Presented at: 2011 IEEE 13th International Conference on e-Health Networking, Applications and Services; June 13-15, 2011; Columbia, MO. [doi: 10.1109/health.2011.6026739]

22. Hasan S, Zhu X, Liu J, Barra CM, Oliveira L, Farri O. Ontology-driven semantic search for Brazilian Portuguese clinical notes. Stud Health Technol Inform 2015;216:1022. [Medline: 26262322]
23. Hill J, Visweswaran S, Ning X, Schleyer TK. Use, impact, weaknesses, and advanced features of search functions for clinical use in electronic health records: a scoping review. Appl Clin Inform 2021 May;12(3):417-428 [FREE Full text] [doi: 10.1055/s-0041-1730033] [Medline: 34261171]
24. mark.js. URL: https://markjs.io/ [accessed 2022-09-27]
25. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004 Jan 01;32(Database issue):D267-D270 [FREE Full text] [doi: 10.1093/nar/gkh061] [Medline: 14681409]
26. List of medical abbreviations. Wikipedia. URL: https://en.wikipedia.org/wiki/List_of_medical_abbreviations [accessed 2022-01-31]
27. Medical Abbreviations & Acronyms. OpenMD. URL: https://openmd.com/dictionary/medical-abbreviations [accessed 2022-02-23]
28. Medical documents. MTSamples. URL: https://www.mtsamples.com/index.asp [accessed 2022-01-31]
29. PassMedicine. URL: https://passmedicine.com/index.php [accessed 2022-01-31]
30. Natarajan K, Stein D, Jain S, Elhadad N. An analysis of clinical queries in an electronic health record search utility. Int J Med Inform 2010 Jul 1;79(7):515-522 [FREE Full text] [doi: 10.1016/j.ijmedinf.2010.03.004] [Medline: 20418155]
31. Chi EA, Chi G, Tsui CT, Jiang Y, Jarr K, Kulkarni CV, et al. Development and validation of an artificial intelligence system to optimize clinician review of patient records. JAMA Netw Open 2021 Jul 01;4(7):e2117391 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.17391] [Medline: 34297075]
32. Schulz S, Daumke P, Fischer P, Müller M, Müller ML. Evaluation of a document search engine in a clinical department system. AMIA Annu Symp Proc 2008 Nov 06:647-651 [FREE Full text] [Medline: 18999064]

Abbreviations

CVA: cerebrovascular accident
EHR: electronic health record
IR: information retrieval
NLP: natural language processing
TIA: transient ischemic attack
UMLS: Unified Medical Language System

Edited by C Lovis; submitted 18.05.22; peer-reviewed by J Hefner; comments to author 15.07.22; revised version received 01.09.22; accepted 07.09.22; published 26.10.22.

Please cite as:
Park EH, Watson HI, Mehendale FV, O'Neil AQ, Clinical Evaluators
Evaluating the Impact on Clinical Task Efficiency of a Natural Language Processing Algorithm for Searching Medical Documents: Prospective Crossover Study
JMIR Med Inform 2022;10(10):e39616
URL: https://medinform.jmir.org/2022/10/e39616
doi:10.2196/39616
PMID:36287591

Original Paper

The Factors Contributing to Physicians' Current Use of and Satisfaction With Electronic Health Records in Kuwait's Public Health Care: Cross-sectional Questionnaire Study

Jawaher Al-Otaibi[1*], MPH; Eleni Tolma[2,3*], MPH, PhD; Walid Alali[1*], PhD; Dari Alhuwail[4*], PhD; Syed Mohamed Aljunid[1*], MD, PhD

[1]Department of Health Policy and Management, College of Public Health, Kuwait University, Kuwait City, Kuwait

[2]Social Behavioral Sciences, College of Public Health, Kuwait Unviersity, Kuwait City, Kuwait

[3]Department of Education, European University Cyprus, Nicosia city, Cyprus

[4]College of Life Sciences, Kuwait University, Kuwait City, Kuwait

[*]all authors contributed equally

Corresponding Author:
Jawaher Al-Otaibi, MPH
Department of Health Policy and Management
College of Public Health
Kuwait University
Sabah Al-Salem University City, Kuwait Ministry of Health Bldg
Medical Record Jamal Abdel Nasser Street, Sulaibikat Club Circle, Al-ASima area
Kuwait City, 12009
Kuwait
Phone: 965 55615073
Email: jawaher309m@gmail.com

Abstract

Background: Electronic health record (EHR) has emerged as a backbone health care organization that aims to integrate health care records and automate clinical workflow. With the adoption of the eHealth care system, health information communication technologies and EHRs are offering significant health care advantages in the form of error reduction, improved communication, and patient satisfaction.

Objective: This study aimed to (1) investigate factors associated with physicians' EHR adoption status and prevalence of EHRs in Kuwait and (2) identify factors predicting physician satisfaction with EHRs in public hospitals in Kuwait.

Methods: This study was conducted at Kuwait's public Al-Jahra hospital from May to September 2019, using quantitative research methods. Primary data were gathered via questionnaires distributed among 295 physicians recruited using convenience sampling. Data were analyzed in SPSS using descriptive, bivariate, and multivariate linear regression, adjusted for demographics.

Results: Results of the study revealed that the controlled variable of gender (β=–.197; *P*=.02) along with explanatory variables, such as training quality (β=.068; *P*=.005), perception of barriers (β=–.107; *P*=.04), and effect on physician (β=.521; *P*<.001) have a significant statistical relationship with physicians' EHR adoption status. Furthermore, findings also suggested that controlled variables of gender (β=–.193; *P*=.02), education (β=–.164; *P*=.03), effect on physician (β=.417; *P*<.001), and level of ease of use (β=.254; *P*<.001) are significant predictors of the degree of physician satisfaction with the EHR system.

Conclusions: The findings of this study had significant managerial and practical implications for creating an inductive environment for the acceptance of EHR systems across a broad spectrum of health care system in Kuwait.

(JMIR Med Inform 2022;10(10):e36313) doi:10.2196/36313

KEYWORDS

health informatics; information systems adoption; electronic health record; EHR; public health informatics

Introduction

Electronic health record (EHR) systems can provide physicians with accurate information to serve patients more efficiently as compared with paper-based systems [1]. A recent literature review indicated that many health care organizations worldwide, especially in low-income countries, still rely on paper-based systems for maintaining patient records [2]. Research suggests that primary issues faced by traditional systems (ie, paper-based systems) are inaccuracy of information, loss of data, and difficulty in sharing information [3]. In Kuwait, many attempts were made to automate clinical workflows in public hospitals. However, lack of organizational readiness and technical knowledge of the user are primary reasons for EHR implementation failure in Kuwait [4].

Evidence of EHR implementation in public and private health care systems suggests that EHRs are more efficient than paper-based electronic record systems [5]. EHRs significantly improve safety, efficiency, and quality of care provided to patients [6,7].

Furthermore, EHR has a significant impact on the performance of health care workers [6]. An EHR system is an integral part of the clinical decision support system, which provides data to a wide range of health care workers and promptly assists in decisions related to diagnosis and treatment, test results, and the cost of health care [7]. Physicians' efficient use of EHR can decrease medical errors and provide every health care professional with accurate and timely information [8]. Health care workers can access information quickly and efficiently through the EHR system, which aids in diagnoses and follow-up treatments [9,10]. EHR covers various types of information, from patient medical history to assimilated information from laboratories, specialists, pharmacists, and insurance companies. The EHR system is not only confined to inpatient care but also extends to aftercare with local general practitioners [11].

In contrast, electronic medical record (EMR) refers to the electronic chart of a patient's medical history assessed by the concerned medical staff. Integration of new technologies, such as Internet of Things, machine learning, artificial intelligence, and decision support, into the electronic health care system module and their implementation has transformed health care. Transformation of traditional data center–based solutions into cloud systems have opened new horizons for applying big data, machine learning, and artificial intelligence [12].

Acceptance of EHR use among physicians in public health care institutes requires considerable investment in training and development. Implementing an EHR system is an issue of change management due to its impact on holistic health care [8]. Thus, the issue of EHR adoption status among physicians has become a significant concern for many public health care institutes, as lack of tech savviness, workflow design, and training are substantial barriers to achieving EHR adoption and satisfaction among physicians [13]

Kuwait provides high-standard health care coverage to its residents. In governmental facilities, free medical treatment is offered to all Kuwaiti nationals. In contrast, foreign residents must pay an annual fee and nominal charges at every visit to access public health care facilities. Kuwait's government spends 4.6% of its gross domestic product on public health expenditures. Kuwait's health care sectors accounted for 11% of the public spending of Kuwait in 2018. There are currently 97 primary health care centers in Kuwait overseen by the Ministry of Health [14].

The history of EMR in Kuwait dates back to 2000, when the Ministry of Health introduced a national EMR system across the entire primary care facilities and hospitals. Moreover, in 2013, a national eHealth strategy was launched that attempted to consolidate all patient health records into a single health record file managed by Kuwait's Ministry of Health and the department of Information Systems [11].

Evidence from a study by Alnashmi et al [15] in Kuwait has shown that most physicians in primary health care settings favor from using the EHR system; however, they suggested additional functionality improvements through digital signatures, integration with artificial intelligence, data warehousing, and big data analytics to enhance the quality of care offered by health care institutes across the country.

Therefore, this research aimed to identify the EHR system adoption status at Al-Jahra hospital in Kuwait, a public hospital operated by the Ministry of Health offering 1234 beds aided by a surgical suite, emergency services, diagnostic center, and outpatient service. It also aimed to measure physicians' satisfaction with the EHR system by using the factors influencing their satisfaction. The study's findings posed significant implications for the public health care system in Kuwait to promote greater use of EHR, which can lead to a decrease in medical errors, better health care services, and overall health care cost reduction [8]. Furthermore, digital transformation is the foremost influential agenda for Kuwait Vision 2035 to strengthen investment in high-quality health care and increase the efficiency of the existing health care system [15].

Recent studies [6,7,9] have focused on the theoretical elements predicting EHR adoption among physicians and satisfaction with the existing EHR system. However, there is limited literature on the theoretical aspects that empirically explain the phenomenon of satisfaction with the use of the EHR system [7]. Therefore, this study mainly aimed to fill the research gap by answering critical questions associated with EHR adoption and the degree of physician satisfaction with the existing EHR system used at Al-Jahra hospital. The study aimed to achieve the following research objectives:

- To investigate the prevalence of EHR and EHR adoption status among physicians at Al-Jahra public hospital.
- To investigate the level of satisfaction with EHR use among physicians working at Al-Jahra hospital.
- To investigate factors predicting physician satisfaction with the EHR system at Al-Jahra hospital.

Figure 1. Perception of training received by physicians at Al-Jahra hospital. EHR: electronic health record.

Methods

Setting

This cross-sectional study occurred at Al-Jahra hospital, a general public hospital in Kuwait. Data were collected from May 2019 to September 2019. The sample size was selected using the Raosoft calculator, which suggests that the sample size for the population of staff members working at Al-Jahra hospital requires 295 responses [16]. The researcher incorporated convenience sampling methods to recruit participants from various stratified groups (ie, gynecology, general physicians, urologists, orthopedics, ENTs, psychiatry, radiology, pathology, cardiology, and gastroenterology).

Theoretical Framework

The Technology Acceptance Model (TAM), an information system theory based on the acceptance and use of technology by people, was used to develop the conceptual framework. TAM provides information on a technology-based framework for understanding the user's adoption of technology and preference for using the advanced technologies, particularly in the workplace environment [6]. The theory is based on the following two primary factors: perceived ease of use and perceived usefulness of technology.

The TAM and the Unified Theory of Acceptance and Use of Technology (UTAUT) are two popular theories used in explaining the use of EHRs; UTAUT helps gauge the degree of physician satisfaction with EHR, as satisfaction is an antecedent of repeated behavioral intention [17].

Survey Assessment Tool

The survey tool was designed and refined, followed by the pilot testing procedure. The questions included in the survey were based on the information extracted from the literature review. The survey was pilot tested among 33 physicians who had experience using the EHR for clarity, readability, and feedback. The instrument questionnaire reported overall reliability of all items, with α=.886, which suggests that the instrument exhibits high internal consistency.

The final version of the survey consisted of 9 sections and 46 items or questions. They were scored on a 5-point Likert response scale ranging from strongly agree to strongly disagree.

The survey was translated into Arabic and then back-translated into English. An expert (DA) also reviewed the survey in Health Information Systems in Kuwait to ensure cultural and contextual fit.

The survey's psychometric properties were established using the Confirmatory Factor Analysis (CFA) [18]. The results of the CFA showed that all items in each construct were retained, with the exception of 2 items in the scale 'Perception of Barriers to Using EHR.' The final instrument consisted of 8 main variables and 35 items. All scales were reliable, with the lowest reliability score of 0.717 (for 'Perception of Barriers to Using EHR') and the highest score of 0.897 (for 'Level of Ease of EHR Function'). There were 2 dependent variables. The first was the physician EHR adoption status, and the second was the degree of physician satisfaction with EHR. The independent variables, as directed by the TAM model, include satisfaction with technical support, preference for using a new EHR system, preference to go back to a paper-based system, perception of barriers to using EHR, the effect of the use of EHR on physician, and level of ease of EHR. The demographics measured in the survey were gender, nationality, age group, education, years of experience, work department, and job title. The quality of the related training was also added as an independent variable based on the theocratical insight of the TAM model, which suggests that training paradigms significantly influence behavioral intentions [19].

Inclusion and Exclusion Criteria

The target population consisted of physicians working at Al-Jahra hospital in Kuwait. Inclusion criteria involved (1) employees of Al-Jahra hospital, (2) physicians, and (3) experience using the EHR system in the hospital, whereas exclusion criteria included (1) former employees of Al-Jahra hospital, (2) administrative staff, (3) nurses, (4) technicians, and (5) physicians working with Al-Jahra on a contractual agreement.

Population and Sampling

According to a previous study, 55% of the physicians in Kuwait are already using an existing adopt EHR system [20]. Considering this adoption rate, a finite population size of 503, a 5% error rate, and a design effect of 1, the required sample consisted of 217 research participants. Assuming a nonresponse rate of 20%, a target sample size of 277 physicians was required for the quantitative study.

Ethics Approval

Ethical approval (2019/1093) was obtained from the Kuwait Ministry of Health Ethical Committee. All research participants signed the informed consent form, which clearly stated the study's purposes, data use, and participants' safety (ie, confidentiality and anonymity).

Statistical Analysis

The paper survey was self-administered. The response rate was 95%. Missing values were treated in SPSS using missing values analysis, which suggested that missing values were completely at random, and there was no pattern that resulted in the pairwise deletion of data.

The data were analyzed using the IBM Statistical Package for Social Sciences (version 23; IBM Corp) [21]. Descriptive statistics analysis was also conducted, followed by bivariate analysis. The most common test used in the bivariate analysis was the Pearson correlation analysis. The final statistical analysis used was multiple regression analysis to test the contribution of the independent variables to the dependent variables (ie, current use of EHR and satisfaction with EHRs), adjusted for the demographics. This was done in two steps; first, in model 1, only demographic variables were added to the analysis; then, in model 2, both demographics and the independent variables were added. The alpha level set for this study was .05.

Results

Descriptive Statistics

Of 295 participants, the majority of the participants were male (n=242, 82%) and non-Kuwaitis, (n=259, 88%) from India, Egypt, Asia, Africa, and other parts of the Middle East and North Africa (or MENA) region. Most of the respondents were generally young (n=120, 40.7%), between 30 and 39 years of age, and were experienced physicians (n=100, 33.9% had 5-10 years of work experience). Most of them were registrars (n=88, 29.8%) and gynecologists (n=114, 38.2%), as shown in Table S1 in Multimedia Appendix 1.

In terms of behavioral characteristics, almost 2 of 5 of the respondents (n=124, 42%) reported using the EHR system for more than 5 years. There was a lack of consensus among respondents regarding the quality of related training received; for example, 89 (30.2%) reported receiving low-quality training, whereas another 89 (30.2%) reported receiving high-quality training on EHR system use.

Bivariate Correlation

Regarding bivariate analysis, the degree of physician satisfaction with the EHR system is strongly correlated with the preference for using the new EHR system (r=0.797) and its effect on physician (r=0.744); it was moderately correlated with satisfaction with technical support (r=0.632) and level of ease of EHR system use (r=0.698), as shown in Tables S2 and S3 in Multimedia Appendix 1.

Multiple Regression Analysis

The first series of multiple regression analyses that included all independent variables in the prediction of the EHR adoption status and adjusted for demographic variables showed that the perception of barriers (β=–.0107; P=.04), the effect of the use of EHR on physician (β=.521; P<.001), and training quality (β=.068; P=.005) are significant predictors of physician EHR adoption status (R^2=0.56), as shown in Table S4 in Multimedia Appendix 1.

In the second series of multiple regression analyses that included all independent variables in the prediction of the degree of satisfaction with EHR use and adjusted for demographic variables, findings showed that gender (β=–.1931; P=.02), education (β=–.164; P=.03), effect on physician (β=.417; P<.001), and level of ease of EHR use (β=.254; P<.001) are significant predictors of the degree of physician satisfaction with the EHR system (R^2=0.62), as shown in Table S5 in Multimedia Appendix 1.

Discussion

Principal Findings

The study's primary purpose was to examine the psychosocial factors associated with physicians' use of EHR and satisfaction with the EHR system at Al-Jahra public hospital in Kuwait. Findings of the study show that the level of EHR adaption status can be predicted with the controlled variable of gender along with explanatory variables, that is, training quality, perception of barriers to using EHR, and effect on the physician. Furthermore, findings also suggested that controlled variables (ie, gender and education) along with explanatory variables (ie, effect on physicians and level of ease of EHR system) significantly influence physician EHR adoption status. The gender of the physician can also play an important role in the use of EHR. In our study, females were more likely than males to use the EHR system and were more satisfied with it, as supported by the literature [22].

The study's findings validate previous studies [12], which highlight the role of risk and trust relationship in predicting EHR adoption status, as findings revealed that the performance and trust relationship implied by the UTAUT model had no impact on physician intention to use an EHR system. This implies that developers, marketers, and medical professionals should improve and optimize patient communication in the EHR system. Our findings validate previous evidence [21] and also suggest that social factors have a negligible effect on physician intention to adopt EHR system, as physicians are driven by their attitudes, ability to control innovation offered by the EHR system, and holistic benefits offered by the system. Findings also validate the role of training in influencing EHR system adoption status among physicians, as evidence from a study by Dunton [23] suggests that training influences perceived usefulness and perceived ease of use as well as external factors, which significantly enhance physician EHR system adoption status.

In terms of the prediction of EHR use, the most important factor was the effect that the use of EHR had on physicians' work. This implies that physicians will be more inclined toward using the EHR system if they perceive a beneficial effect of the use of EHR on their work. In addition, the length of use of EHR

also had a positive contribution to the prediction of EHR use. This is not surprising, since using the EHR system for an extended period will lead to adopting the EHR system, according to a study by Liang et al [20].

Regarding the prediction of physicians' satisfaction with the use of EHR, the most significant contributor was the effect of EHR use on physicians' work, as supported by the findings of a previous study [24]. Specifically, it was found that the higher the perceptions of the positive effects of EHR on physicians' work, the more likely it will be for the physicians to be satisfied with the use of the EHR system. The second most important contributor was the degree of ease of EHR use. Consistent with the findings of other studies, as physicians start to experience the ease of using the EHR system, they will start adopting the EHR system [3]. Moreover, another study [25] found that perceived usefulness and perceived ease of use increase the acceptance of using the EHR system and hence the satisfaction with it. As it was explained, accepting the use of the EHR system was an indication of the level of satisfaction of the physicians. Therefore, it was not surprising that this study found the degree of ease of using EHR as an important factor in satisfaction with it. In other words, as physicians perceived the EHR system to be easy to use, they were more likely to use it and experience higher satisfaction levels.

Another unique finding concerning the satisfaction with EHR was related to the academic background of participants. According to the results, this characteristic was a significant factor. This implies that the knowledge and academic experience of the physicians might have an impact on their satisfaction level. Those physicians with higher qualifications will tend to be more satisfied with the EHR system compared with others because they might believe that using the EHR system would allow them to serve the patients better. Evidence from a study [26] demonstrated that physicians with higher levels of education had higher levels of satisfaction with EHR use.

Finally, age was also significant in the prediction of EHR use but in a negative way. Older physicians were more reluctant to use the EHR system, as supported by a previous study [17]. In our study, the demographics were treated as controlled variables. Therefore, their effect on the regression model and other variables was neglected.

From a programmatic perspective, the following are some recommendations for public health professionals in their effort to promote the use of EHR and increase satisfaction with EHR among those who are already using it in governmental hospitals of Kuwait:

- Professionals should first conduct a needs assessment, identify perceived barriers among physicians, and try to address those barriers.
- Public health professionals should focus on improving the functionality of the EHR system and make it as easy as possible to operate; this will encourage physicians to use it more often and rely on the EHR system when seeing patients.
- Public health professionals are advised to emphasize promoting the EHR system's positive effects on physicians' work, which could be done through health communication campaigns.

Limitations

There were some limitations to the study. The results of this study were only limited to Al-Jahra hospital, where the study took place. However, this study can be generalized, and the outcomes can be easily applied to other public sector hospitals in Kuwait, as the research examined satisfaction with EHR and adoption of the EHR system. Second, the TAM was used to assess the adoption of EHR and satisfaction with it. However, the TAM and UTUAT models were not fully applied to this study, as the attitude was not examined. Due to the questionnaire being long, the decision was made to shorten it, so more physicians were interested in filling out the questionnaire and participating in the study. The study did not use purposive sampling as convenience sampling for data collection. The use of a random sample might have produced a different result. The findings of the study are not generalizable to all hospitals in Kuwait. Thus, more studies need to be conducted to validate whether other public hospitals exhibit the same phenomenon.

Future Research

Since this study could not cover all aspects that might be useful in examining the satisfaction with EHR system and current use of it, the following future studies must be carried out. First, the theoretical framework should be expanded to include physicians' attitudes toward using EHR. Research should be conducted that fully uses the TAM by including physicians' attitudes. It might prove important to examine physicians' attitudes, since some physicians might have a positive predisposition toward using EHR but still not use it. It would be interesting from a theoretical and programmatic perspective to examine how attitude relates to intention by itself. Second, a follow-up qualitative study through several interviews with senior physicians and hospital officials should be conducted. Such a study will help identify more in-depth information behind using or not using the EHR system. For instance, qualitative research can complement quantitative research results and help us discover the perceived barriers to adopting the EHR system. In addition, qualitative research can help us find answers to surprising results, such as the fact that women are more likely to adopt an EHR system. This study can be replicated in other governmental hospitals in Kuwait to reach a better understanding of how prevalent the use of EHR is and the degree of satisfaction with its use.

Conclusions

There are important takeaways from the results of this study. First, there is still a need to further expand the EHR system adoption at Al-Jahra hospital, since almost 1 in 5 physicians has never used EHR or has used EHR for less than a year. This could be justified as they may have joined the hospital recently. Second, to increase the adoption rate and satisfaction with the current use of EHR among physicians, public health professionals can make the benefits of EHR adoption more visible to the physicians, remove perceived barriers, make the use of the EHR system as easy as possible, and incorporate a high-quality related training, while providing continuous technical support. Results from this study can be helpful to other

governmental hospitals in Kuwait in their efforts to enhance the levels of adoption and satisfaction with the EHR system. The EHR system has many benefits, and it can be fully realized only when all physicians in governmental hospitals in Kuwait fully adopt it.

Acknowledgments

The authors would like to thank Kuwait University, the Ministry of Health, and Al-Jahra Hospital for facilitating this research. This study did not receive any funding.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Tables S1-S5.
[DOCX File , 32 KB - medinform_v10i10e36313_app1.docx]

References

1. Williams DC, Warren RW, Ebeling M, Andrews AL, Teufel Ii RJ. Physician use of electronic health records: survey study assessing factors associated with provider reported satisfaction and perceived patient impact. JMIR Med Inform 2019 Apr 04;7(2):e10949 [FREE Full text] [doi: 10.2196/10949] [Medline: 30946023]
2. Churi P, Pawar A, Moreno-Guerrero A. A comprehensive survey on data utility and privacy: taking Indian healthcare system as a potential case study. Inventions 2021 Jun 23;6(3):45. [doi: 10.3390/inventions6030045]
3. Denhere PT, Zhou M, Ruhode E. A practitioner centred assessment on health information systems interoperability readiness in Zimbabwe: mixed study approach. ujph 2022 Feb;10(1):1-14. [doi: 10.13189/ujph.2022.100101]
4. Bisrat A, Minda D, Assamnew B, Abebe B, Abegaz T. Implementation challenges and perception of care providers on electronic medical records at St. Paul's and Ayder Hospitals, Ethiopia. BMC Med Inform Decis Mak 2021 Nov 02;21(1):306 [FREE Full text] [doi: 10.1186/s12911-021-01670-z] [Medline: 34727948]
5. Keshta I, Odeh A. Security and privacy of electronic health records: concerns and challenges. Egypt Inform J 2021 Jul;22(2):177-183. [doi: 10.1016/j.eij.2020.07.003]
6. Almutairi BA, Potts HWW, Al-Azmi SF. Physicians' perceptions of electronic prescribing with electronic medical records in Kuwaiti primary healthcare centres. Sultan Qaboos Univ Med J 2018 Nov 28;18(4):e476-e482 [FREE Full text] [doi: 10.18295/squmj.2018.18.04.008] [Medline: 30988966]
7. Hossain A, Quaresma R, Rahman H. Investigating factors influencing the physicians' adoption of electronic health record (EHR) in healthcare system of Bangladesh: an empirical study. Int J Inf Manage 2019 Feb;44:76-87. [doi: 10.1016/j.ijinfomgt.2018.09.016]
8. Agbese FAO, Ikonne SN. Electronic health records (EHRs) use and quality healthcare delivery by physicians in tertiary hospitals in federal capital territory, Nigeria. LPP (e-journal) 2021:1-12 [FREE Full text]
9. Alshahrani A, Stewart D, MacLure K. A systematic review of the adoption and acceptance of eHealth in Saudi Arabia: views of multiple stakeholders. Int J Med Inform 2019 Aug;128:7-17. [doi: 10.1016/j.ijmedinf.2019.05.007] [Medline: 31160014]
10. Haried P, Claybaugh C, Dai H. Evaluation of health information systems research in information systems research: a meta-analysis. Health Informatics J 2017 Apr 25;25(1):186-202. [doi: 10.1177/1460458217704259]
11. Alaslawi H, Berrou I, Alhuwail D, Aslanpour Z. Status and trends of e-Health tools in Kuwait: a narrative review. JHIDC 2019;13(2):1 [FREE Full text]
12. Arfi WB, Nasr IB, Kondrateva G, Hikkerova L. The role of trust in intention to use the IoT in eHealth: application of the modified UTAUT in a consumer context. Technol Forecast Soc Change 2021 Jun;167:120688. [doi: 10.1016/j.techfore.2021.120688]
13. Elharish SF, Denna I, Abdelsalam MM, Elberkawi EK. International Conference on Data Science, E-learning and Information Systems 2021. 2021 Presented at: DATA'21; April 5-7; New York, NY p. 40-46. [doi: 10.1145/3460620.3460628]
14. Besa G. [Left ventricular aneurysmectomy: technical evolution and results]. G Ital Cardiol 1992 Mar;22(3):331-336. [Medline: 1426774]
15. Alnashmi M, Salman A, AlHumaidi H, Yunis M, Al-Enezi N. Exploring the health information management system of Kuwait: lessons and opportunities. ASI 2022 Feb 15;5(1):25. [doi: 10.3390/asi5010025]
16. Sample size calculator. Raosoft. 2004. URL: http://www.raosoft.com/samplesize.html [accessed 2022-09-19]
17. Aldosari B, Al-Mansour S, Aldosari H, Alanazi A. Assessment of factors influencing nurses acceptance of electronic medical record in a Saudi Arabia hospital. IMU 2018;10:82-88. [doi: 10.1016/j.imu.2017.12.007]
18. Pallant J. SPSS survival manual: a step-by-step guide to data analysis using SPSS. London, UK: Routledge; 2020:1-368.

19. Pai MMM, Ganija R, Pai RM, Sinha RK. Standard electronic health record (EHR) framework for Indian healthcare system. Health Serv Outcomes Res Methodol 2021;21:339-362. [doi: 10.1007/s10742-020-00238-0]
20. Liang J, Li Y, Zhang Z, Shen D, Xu J, Zheng X, et al. Adoption of electronic health records (EHRs) in China during the past 10 years: consecutive survey data analysis and comparison of Sino-American challenges and experiences. J Med Internet Res 2021 Feb 18;23(2):e24813 [FREE Full text] [doi: 10.2196/24813] [Medline: 33599615]
21. El-Yafouri R, Klieb L, Sabatier V. Psychological, social and technical factors influencing electronic medical records systems adoption by United States physicians: a systematic model. Health Res Policy Syst 2022 May 02;20(1):48 [FREE Full text] [doi: 10.1186/s12961-022-00851-0] [Medline: 35501897]
22. Mohammed HT, Hyseni L, Bui V, Gerritsen B, Fuller K, Sung J, et al. Exploring the use and challenges of implementing virtual visits during COVID-19 in primary care and lessons for sustained use. PLoS One 2021 Jun 24;16(6):e0253665 [FREE Full text] [doi: 10.1371/journal.pone.0253665] [Medline: 34166441]
23. Dunton S. Assessment of EHR implementation and training processes at a pilot site for a national initiative: a program evaluation. College of Nursing in Spokane, Washington State University. 2021. URL: https://s3.wp.wsu.edu/uploads/sites/3014/2021/11/559-Abstract-Dunton.pdf [accessed 2022-09-27]
24. Chang S, Lu M, Pan T, Chen C. Evaluating the e-Health cloud computing systems adoption in Taiwan's healthcare industry. Life (Basel) 2021 Apr 02;11(4):310 [FREE Full text] [doi: 10.3390/life11040310] [Medline: 33918246]
25. Spatar D, Kok O, Basoglu N, Daim T. Adoption factors of electronic health record systems. Technol Soc 2019 Aug;58:101144. [doi: 10.1016/j.techsoc.2019.101144]
26. Topaz M, Ronquillo C, Peltonen L, Pruinelli L, Sarmiento RF, Badger MK, et al. Advancing nursing informatics in the next decade: recommendations from an international survey. Stud Health Technol Inform 2016;225:123-127. [Medline: 27332175]

Abbreviations

EHR: electronic health record
EMR: electronic medical record
TAM: Technology Acceptance Model
UTAUT: Unified Theory of Acceptance and Use of Technology

Edited by C Lovis; submitted 10.01.22; peer-reviewed by W Zhang, Y Chu, KM Kuo; comments to author 11.04.22; revised version received 02.08.22; accepted 07.09.22; published 07.10.22.

Please cite as:
Al-Otaibi J, Tolma E, Alali W, Alhuwail D, Aljunid SM
The Factors Contributing to Physicians' Current Use of and Satisfaction With Electronic Health Records in Kuwait's Public Health Care: Cross-sectional Questionnaire Study
JMIR Med Inform 2022;10(10):e36313
URL: https://medinform.jmir.org/2022/10/e36313
doi:10.2196/36313
PMID:36206039

Original Paper

Successful Integration of EN/ISO 13606–Standardized Extracts From a Patient Mobile App Into an Electronic Health Record: Description of a Methodology

Santiago Frid[1,2], MSc, MD; Maria Angeles Fuentes Expósito[3], MSc; Inmaculada Grau-Corral[3,4], PhD; Clara Amat-Fernandez[5], MSc; Montserrat Muñoz Mateu[2,6], MD, PhD; Xavier Pastor Duran[1,2], MD, PhD; Raimundo Lozano-Rubí[1,2], MD, PhD

[1]Medical Informatics Unit, Hospital Clínic de Barcelona, Barcelona, Spain

[2]Universitat de Barcelona, Barcelona, Spain

[3]FundiSYS, Barcelona, Spain

[4]mHealth Observatory, Hospital Clínic de Barcelona, Barcelona, Spain

[5]Fundació Clínic per a la Recerca Biomèdica, Barcelona, Spain

[6]Oncology Unit, Hospital Clínic de Barcelona, Barcelona, Spain

Corresponding Author:
Santiago Frid, MSc, MD
Medical Informatics Unit
Hospital Clínic de Barcelona
Villarroel 170
Barcelona, 08036
Spain
Phone: 34 932 27 54 00 ext 3344
Email: santifrid@gmail.com

Abstract

Background: There is an increasing need to integrate patient-generated health data (PGHD) into health information systems (HISs). The use of health information standards based on the dual model allows the achievement of semantic interoperability among systems. Although there is evidence in the use of the Substitutable Medical Applications and Reusable Technologies on Fast Healthcare Interoperability Resources (SMART on FHIR) framework for standardized communication between mobile apps and electronic health records (EHRs), the use of European Norm/International Organization for Standardization (EN/ISO) 13606 has not been explored yet, despite some advantages over FHIR in terms of modeling and formalization of clinical knowledge, as well as flexibility in the creation of new concepts.

Objective: This study aims to design and implement a methodology based on the dual-model paradigm to communicate clinical information between a patient mobile app (Xemio Research) and an institutional ontology-based clinical repository (OntoCR) without loss of meaning.

Methods: This paper is framed within Artificial intelligence Supporting CAncer Patients across Europe (ASCAPE), a project that aims to use artificial intelligence (AI)/machine learning (ML) mechanisms to support cancer patients' health status and quality of life (QoL). First, the variables "side effect" and "daily steps" were defined and represented with EN/ISO 13606 archetypes. Next, ontologies that model archetyped concepts and map them to the standard were created and uploaded to OntoCR, where they were ready to receive instantiated patient data. Xemio Research used a conversion module in the ASCAPE Local Edge to transform data entered into the app to create EN/ISO 13606 extracts, which were sent to an Application Programming Interface (API) in OntoCR that maps each element in the normalized XML files to its corresponding location in the ontology. This way, instantiated data of patients are stored in the clinical repository.

Results: Between December 22, 2020, and April 4, 2022, 1100 extracts of 47 patients were successfully communicated (234/1100, 21.3%, extracts of side effects and 866/1100, 78.7%, extracts of daily activity). Furthermore, the creation of EN/ISO 13606–standardized archetypes allows the reuse of clinical information regarding daily activity and side effects, while with the creation of ontologies, we extended the knowledge representation of our clinical repository.

Conclusions: Health information interoperability is one of the requirements for continuity of health care. The dual model allows the separation of knowledge and information in HISs. EN/ISO 13606 was chosen for this project because of the operational mechanisms it offers for data exchange, as well as its flexibility for modeling knowledge and creating new concepts. To the best of our knowledge, this is the first experience reported in the literature of effective communication of EN/ISO 13606 EHR extracts between a patient mobile app and an institutional clinical repository using a scalable standard-agnostic methodology that can be applied to other projects, data sources, and institutions.

(JMIR Med Inform 2022;10(10):e40344) doi:10.2196/40344

KEYWORDS

health information interoperability; mobile app; health information standards; artificial intelligence; electronic health records; machine learning

Introduction

Importance of Patient-Generated Health Data

Traditionally, physicians were the only actor who registered patient data in health information systems (HISs). In recent years, the focus has shifted toward more active participation by patients in their own health care, particularly by means of patient-generated health data (PGHD) [1].

One relevant source of PGHD are wearables, electronic devices that connect to the body surface of patients and can transmit data regarding many biological variables. The number of such devices that generate valuable data is growing considerably.

Furthermore, patient experience has been progressively incorporated into health care processes with the objective to optimize them. One of the most relevant measures of outcomes is the patient-reported outcome measures (PROMs), which record patients' perception of disease, including relevant symptoms and emotional distress [2]. In the context of the increasingly adopted value-based health care model, Michael Porter developed a formula: value = (results that matter to the patient)/costs [3,4]. In this model, it is key that patients report the results that matter most to them using indicators provided by PROMs [5].

Increasingly, all of these data come from patient mobile apps, and they need to be integrated into HISs for their use in the caregiving process (primary use) or for research purposes (secondary use). However, given the large number of HISs that coexist even within a single health organization, this proves to be highly challenging.

Interoperability in Health Information Systems

To share clinical information in such a way that it can be unequivocally interpreted, both syntactically and semantically, by 2 or more systems, a common health information standard must be used.

European Norm/International Organization for Standardization (EN/ISO) 13606 is a health information standard that seeks to define a rigorous and stable architecture for communicating health records of a single patient, preserving the original clinical meaning. It is based on a dual model proposed by OpenEHR [6] that includes a reference model (with the necessary components, and their constraints, to represent electronic health record [EHR] extracts) and an archetype model (for the formalization of the clinical domain concepts according to the reference model) [7,8]. Thus, EN/ISO 13606 was designed for the exchange of EHR extracts with full meaning and a high compatibility with OpenEHR [9].

The Fast Healthcare Interoperability Resources (FHIR) standard was developed by Health Level 7 (HL7) with the intention to use modern communication standards for the agile creation of health data communication infrastructures [10]. FHIR's 80/20 rule (focus on 20% of the requirements that satisfy 80% of the interoperability needs) centers on simplicity rather than completeness. FHIR also provides a health information standard to Substitutable Medical Applications and Reusable Technologies (SMART), a framework that enables medical apps to be written once and run unmodified across different health care information technology (IT) systems [11].

EN/ISO 13606's advantages over FHIR in terms of modeling and formalization of clinical knowledge, as well as flexibility in the creation of new concepts, suggest it could play a role in the communication of EHR extracts with mobile apps, despite the limited existing evidence. This could be particularly useful in complex scenarios of health data exchange between nodes [12].

The ASCAPE Project

This paper is framed within the Artificial intelligence Supporting CAncer Patients across Europe (ASCAPE) project, where breast and prostate cancer, 2 of the most prevalent types of cancer, are considered [13]. One of the main purposes of the project is to use powerful artificial intelligence (AI)/machine learning (ML) mechanisms to support cancer patients' health status and quality of life (QoL) in 4 different pilots [14,15].

Within the ASCAPE project, clinical partners identified previously validated questionnaires used to capture different QoL issues for both types of cancer. AI-based models ingest data from such questionnaires, as well as data regarding daily activity, side effects, and physicians' interventions, to predict and suggest improvements in patient QoL issues. Hence, ASCAPE prospectively investigates an AI-based approach toward a personalized follow-up strategy for cancer patients focusing on their QoL issues.

The approach chosen in the project to properly process sensitive medical data is federated learning (FL), a decentralized ML technique where local data are used to train shared global models with a central server, keeping the sensitive data locally.

Objectives

The aim of this study was to design and implement a methodology based on the dual model paradigm in order to communicate clinical information between a patient mobile app and an institutional clinical repository, without loss of meaning. This implies a series of specific objectives:

- To conceptually represent information regarding daily activity and side effects by means of ontologies
- To define a set of archetypes based on EN/ISO 13606 for the standardization and consolidation of patient data in clinical repositories
- To create a scalable conversion module for mobile apps, within the Hospital Clínic de Barcelona's (HCB) environment, to transform local data and generate EN/ISO 13606–compliant EHR extracts
- To validate the methodology through the successful generation and integration of EHR extracts sent from Xemio Research, a patient mobile app, into the institutional ontology-based clinical repository, OntoCR.

Methods

Ethical Considerations

This study was approved by the Hospital Clínic de Barcelona Ethics Committee for Investigation with Drugs (HCB/2020/0971).

Systems and Servers

OntoCR

Traditionally, HISs were developed with a focus on financial and administrative activities, whereas clinical data have been merely translated from paper records to electronic databases. Clinical concepts and the relationships between them have been poorly developed.

OntoCR is an ontology-driven clinical repository conforming to the EN/ISO 13606 standard that uses ontologies for different purposes [16,17]. On the one hand, they define a conceptual architecture centered on the representation of the clinical process and clinical knowledge. By representing a metamodel of health information standards, classifications, and terminologies, OntoCR can also achieve syntactic and semantic interoperability between different HISs. On the other hand, OntoCR uses an ontology that defines the available elements that can be used to build an app. These elements are used by portlets to create a graphical user interface (GUI) deployed in Liferay [18], thus allowing users to access, visualize, enter, and modify structured data through a web-based clinical workstation. OntoCR is linked to the HCB's EHRs (SAP) using the patient ID, and it can be accessed via SAP or its own website.

Xemio Research

Xemio Research was developed for breast cancer patients, providing them with proper information, allowing the tracking of symptoms, and collecting physical activity data from its users on a daily basis (steps, time of activity, and calories). The deployment of Xemio Research's backend takes place within the gated area of the HCB, with a dedicated server (CentOS Linux) whose database is modeled object-oriented in PostgresDB without normalized codes for secondary effect or activity references, just literals names in Spanish.

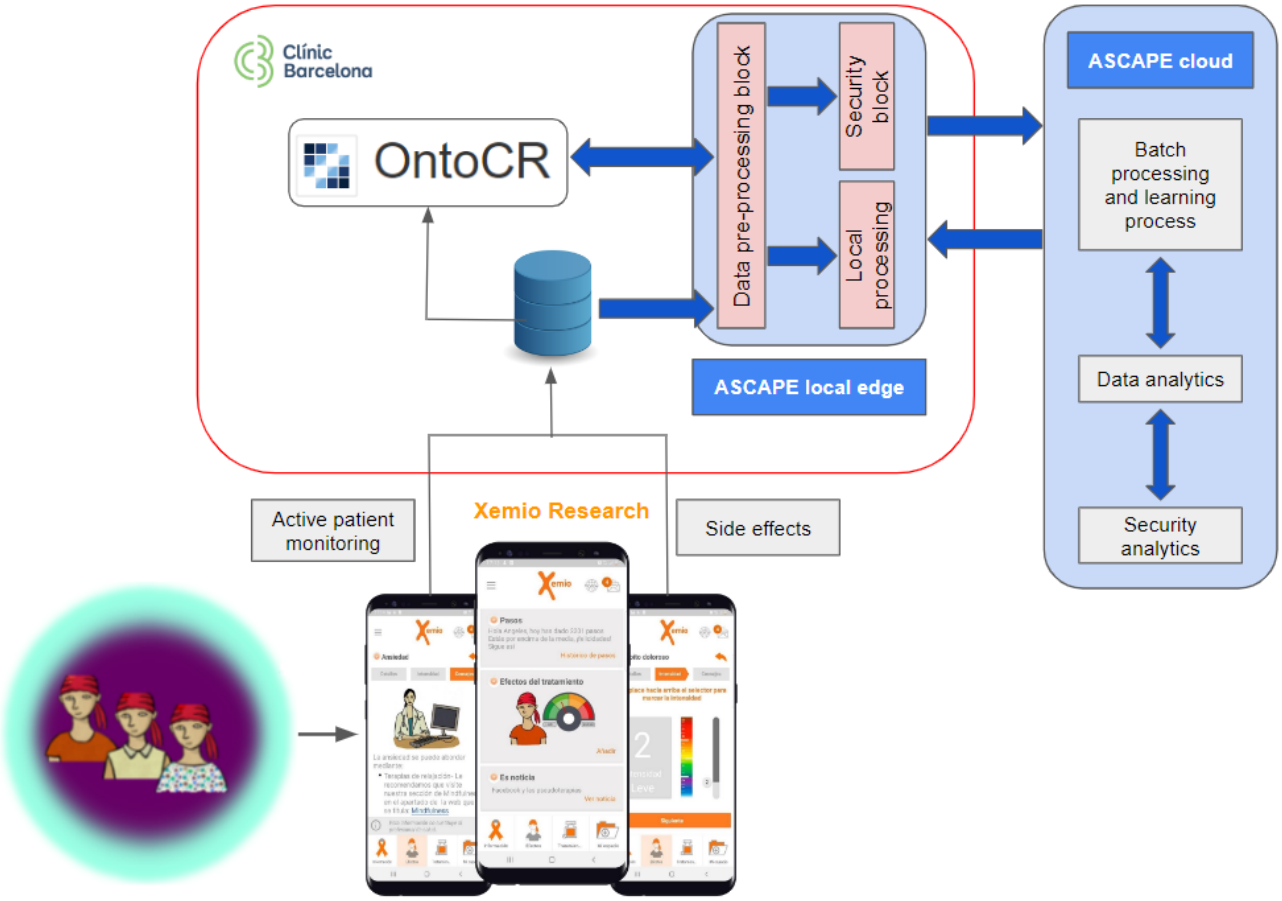
Xemio Research is published in Apple App Store and Google Play Store, with access restricted to study participants. The app was installed on the patient's phone by the field researcher during the first visit, where the patient provided signed informed consent. This generated a Xemio Research ID, which was then registered and linked to the ASCAPE ID in OntoCR by the field researcher.

ASCAPE Local Edge

Due to the sensitive nature of real patient data and the security and data treatment requirements of the project, the ASCAPE architecture was implemented in a dedicated server (ASCAPE Local Edge) within the HCB's environment, supervised by the local IT department (Figure 1).

This architecture was deployed using Kubernetes (k8s) [19], an open source software that accelerates the implementation and administration of containers on a large scale. These containers maintain the microservices needed for the functioning of the project; the processes of data extraction, transformation, and load (ETL); the normalization of retrospective data provided by the HCB; patient anonymization; and predictions offered by AI models. The aforementioned normalization of local data is performed by identifying variables of interest and transforming them to the ASCAPE Common Data Model and HL7 FHIR [15], thus generating a uniform ASCAPE-standardized database for training data sets to feed the AI engines. Furthermore, Local Edge generates and updates ASCAPE's AI predictive models [14], which are shared and evaluated in its accuracy in the federated node.

Figure 1. Information systems within the ASCAPE project. Patients register side effects in Xemio Research, which also tracks patients' daily steps. These data are standardized using a conversion module within the HCB environment (see the Methodology section, step 3), and it is both stored in the OntoCR and sent to ASCAPE Local Edge, which generates and updates ASCAPE's AI predictive models, which are shared and evaluated in its accuracy in the federated node. AI: artificial intelligence; ASCAPE: Artificial intelligence Supporting CAncer Patients across Europe; HCB: Hospital Clínic de Barcelona.



Methodology

The methodology comprises a series of steps to achieve successful sharing of standardized clinical information between a patient mobile app and an institutional clinical repository.

Step 1: Definition of Variables to Communicate and Creation of EN/ISO 13606 Archetypes

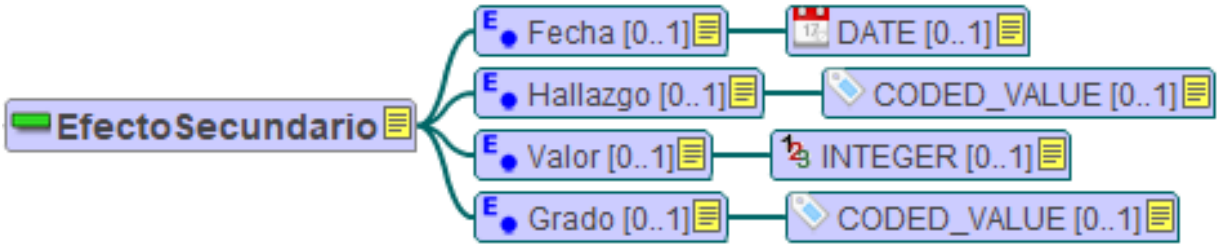
The first step in the methodology is to define clinical variables that need to be communicated through EHR extracts. Since this study was framed within the ASCAPE project, we identified variables that needed to be registered and could be recorded with Xemio Research:

- Daily activity: date, steps, calories, and duration
- Side effects: date, finding, value, and severity

To share information standardized with EN/ISO 13606, archetypes that define the chosen variables must be created. EN/ISO 13606's reference model has multiple components, including the *entry* ("a result of one clinical action, one observation, one clinical interpretation, or one intention") and its *elements* ("The leaf node of the EHR hierarchy, containing a single data value").

Figure 2 shows a mindmap created with the LinkEHR tool [20] of the "side effect" entry archetype. Data types used are those established by the reference model.

Figure 2. Mindmap of the "side effect" archetype (in Spanish), edited with LinkEHR. The "side effect" entry has 4 elements: date, finding, value, and severity.



Step 2: Creation of Ontologies

Once the archetypes are generated, the clinical concepts defined by them must be represented in both systems (mobile app and clinical repository). The functionalities needed to record these variables had already been developed in Xemio Research. For OntoCR, the Medical Informatics Unit at the HCB created the corresponding ontologies to represent these concepts.

A locally developed ontology named Ontoclinic already had a representation of most of the clinical findings that would be used for this project. Hence, the remaining concepts were modeled and added to Ontoclinic, which was later imported into the ASCAPE ontologies. Ontoclinic also includes metaclasses that represent standard classifications and terminologies. Thus, by indicating that a given class is an instance of the Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT) metaclass, it allows the normalization of concepts (see Figure 3). Both *finding* and *severity* were coded with the international edition of SNOMED CT using this approach.

Afterward, both Xemio Research and OntoCR had to model local concepts following the standard. In the first case, this was performed by a conversion module in Local Edge, independent from the app. This component is configured by a text document in JSON format that contains the SNOMED CT codes for each side effect and its severity. The procedure was developed in Python, and it transforms, conceptualizes, and generates daily EN/ISO 13606 EHR extracts with the data of Xemio Research users.

In OntoCR, the modeling was performed by means of ontologies. The HCB Medical Informatics Unit created an ontology that incorporates both EN/ISO 13606 reference and archetype models, enabling the capability of representing clinical data that conform to the standard. Therefore, new ontologies of each entry were created, where the concepts defined in the archetypes were mapped to the EN/ISO 13606 structure.

Figure 3 shows the ontological modeling of concepts described in steps 1 and 2. The upper-left image displays the Secondary_effect class of the Ontoclinic ontology, with its properties *date*, *severity*, *finding*, and *value*. The lower-left image shows the modeling of the Ontoclinic Severe class with SNOMED CT, which was performed by making the concept an instance of the SCT metaclass, thus allowing its binding to a code Uniform Resource Identifier (URI) and a concept ID. Finally, the right image displays the Secondary_effect class modeled with EN/ISO 13606 as a subclass of EN/ISO 13606 ENTRY, therefore inheriting properties of its superclass. Once the ontologies that represent the clinical concepts are created, they are uploaded to OntoCR (Figure 4), where they will be ready to receive instantiated patient data.

Figure 3. Ontologies of "side effect" modeled locally (upper left) and with EN/ISO 13606 (right) and modeling of the concept "severe" using the international edition of SNOMED CT (lower left), all of them in Spanish and edited with Protégé. EN/ISO: European Norm/International Organization for Standardization; SNOMED CT: Systematized Nomenclature of Medicine – Clinical Terms.

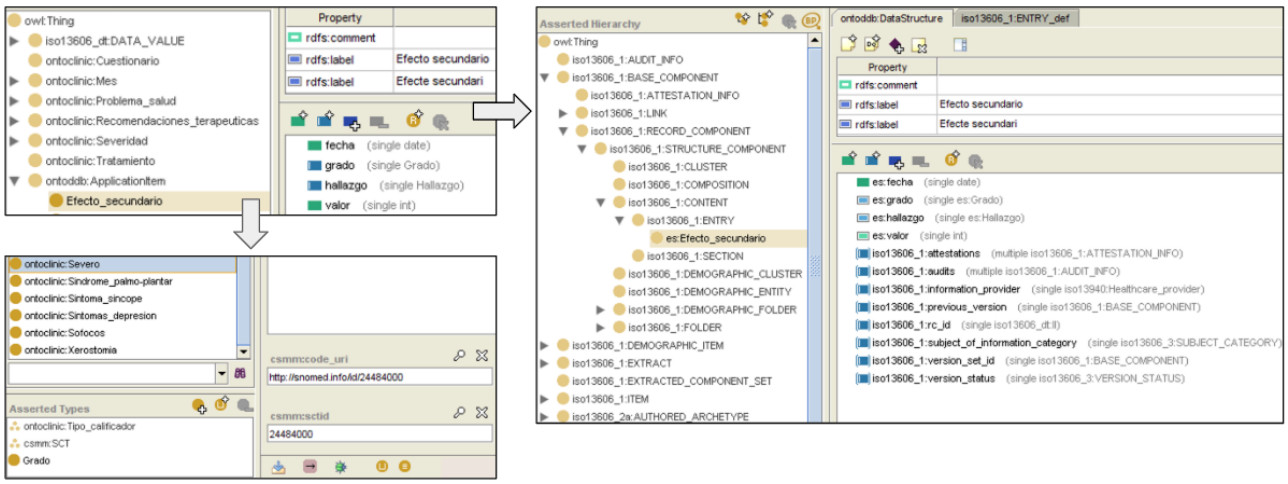
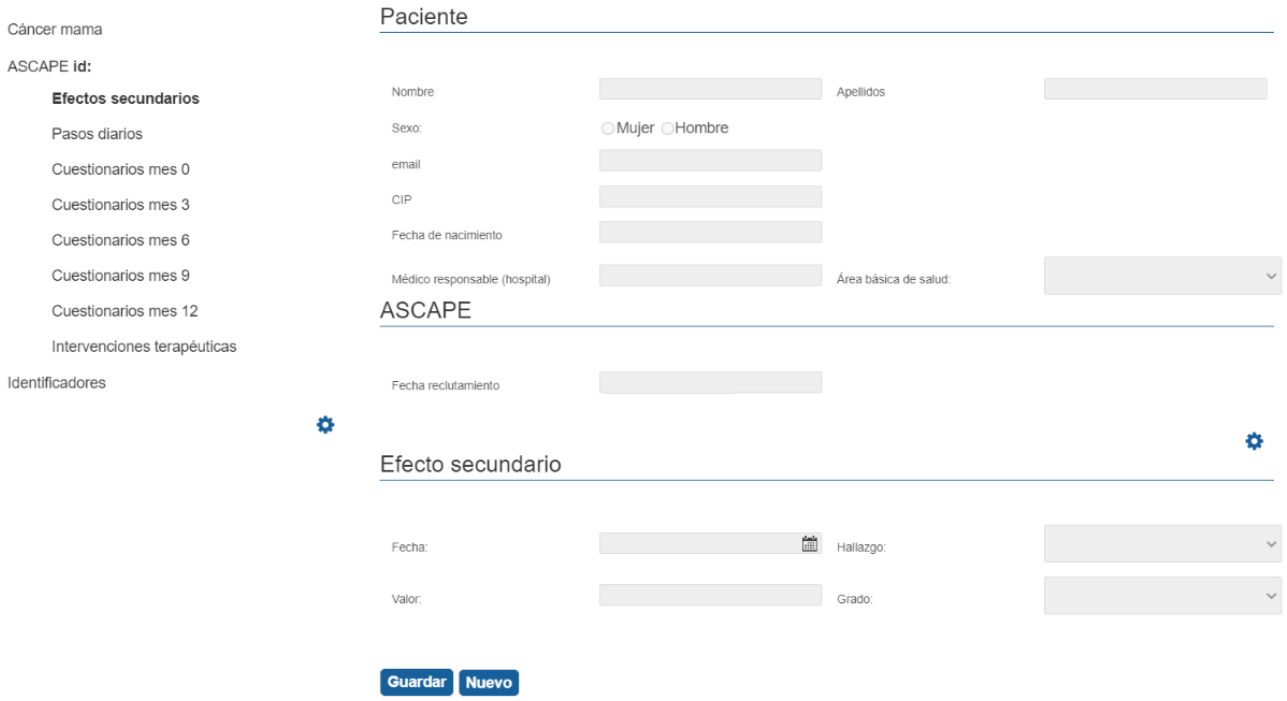


Figure 4. OntoCR GUI for physicians. The ontology modeling the clinical variables is visualized as a web-based structured form. The side effects menu item is selected in the hierarchical menu on the left side of the image. The right side of the image shows the properties regarding patient information, ASCAPE recruitment date, and side effects. ASCAPE: Artificial intelligence Supporting CAncer Patients across Europe; GUI: graphical user interface.



Step 3: Communication of Standardized Extracts

After the variables were defined, represented, and standardized in both systems, extracts were ready to be communicated. Xemio Research has integrated services that transmit extracts with pseudo-anonymized data of either side effects or daily activity collected by the app to an Application Programming Interface (API) in OntoCR, which allows the insertion of extracts into the ontology. This way, instantiated data of patients are stored in OntoCR.

Regarding data security and privacy, Xemio Research generated extracts with anonymous identifiers that were assigned to the patients during recruitment. OntoCR stores the information of both Xemio Research IDs and ASCAPE IDs, so it can integrate the data from the extracts with the rest of the clinical records. Therefore, there is no need for the app to receive data from the hospital's HIS, which is why communication between Xemio Research and OntoCR is unidirectional. This ensures the confidentiality of the real patient data that are managed.

An example of an EN/ISO 13606 EHR extract of side effects is displayed in Figure 5, where the "Wakefulness" finding (coded with the SNOMED CT concept ID 365930002) is recorded.

Figure 6 shows an overview of the process of knowledge modeling and extract communication between Xemio Research and OntoCR. Archetypes created with LinkEHR based on clinical concepts are used as templates to model knowledge in ontologies using Protégé. The addition of ontological layers that contain the metamodels of terminologies, such as SNOMED CT, and health information standards, such as EN/ISO 13606, allow for semantic interoperability of the information. These ontologies, without instantiated data yet, are uploaded to OntoCR.

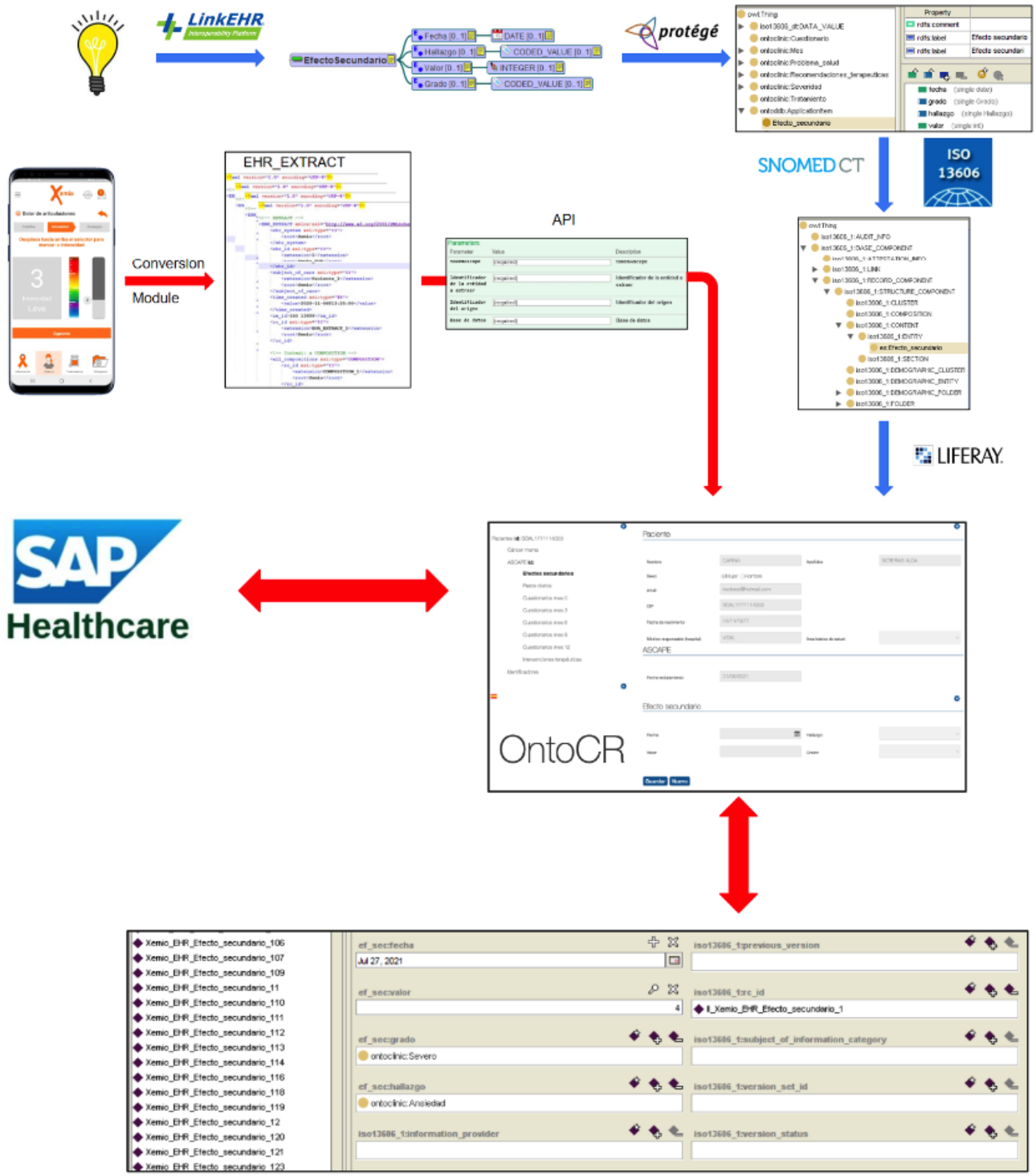
Figure 5. Example of a deidentified EHR extract of side effects. EHR: electronic health record.

```
<content xsi:type="ENTRY">
    <archetype_id xsi:type="II">
        <extension>at0000</extension>
        <root>ISO-EN13606-ENTRY.EfectoSecundario.v2</root>
        <identifier_name>Efecto Secundario</identifier_name>
    </archetype_id>

    <rc_id xsi:type="II">
        <extension>Efecto_secundario_123</extension>
        <root>Xemio_EHR</root>
<identifier_name>Xemio_EHR_Efecto_secundario_123</identifier_name>
    </rc_id>
    <items xsi:type="ELEMENT">
        <archetype_id xsi:type="II">
            <extension>at0001</extension>
            <root>ISO-EN13606-ENTRY.EfectoSecundario.v2</root>
            <identifier_name>Fecha</identifier_name>
        </archetype_id>
      <rc_id xsi:type="II">
            <extension>fecha_123</extension>
            <root>Xemio_EHR</root>
        </rc_id>
        <value xsi:type="DATE">
            <value>2021-10-27</value>
        </value>
</items>

<items xsi:type="ELEMENT">
    <archetype_id xsi:type="II">
        <extension>at0005</extension>
        <root>ISO-EN13606-ENTRY.EfectoSecundario.v2</root>
        <identifier_name>Hallazgo</identifier_name>
    </archetype_id>

  <rc_id xsi:type="II">
        <extension>Hallazgo_123</extension>
        <root>Xemio_EHR</root>
    </rc_id>
    <value xsi:type="CV">
        <code>365930002</code>
        <code_system xsi:type="OID">
            <oid>2.16.840.1.113883.6.96</oid>
        </code_system>
    </value>
</items>
```

Patients enter information on Xemio Research, which normalizes it through a conversion module, thus creating EN/ISO 13606 EHR extracts. These extracts are sent to the API of OntoCR, which inserts patient data into the ontology. The lower image displays a list of instances of side effects, with the corresponding values of the properties *date*, *value*, *severity*, and *finding* entered in Xemio Research by the patient. Furthermore, an instance of the rc_id EN/ISO 13606 property was inserted, indicating the unique identifier by which this instance is referenced in the EHR system.

The process for developing the communication of extracts started in November 2020 and finished in March 2022, with effective deployment in a production environment. On March 14, 2022, all EHR extracts corresponding to retrospective data were sent, and thereafter, extracts were sent daily.

Figure 6. Overview of the process of knowledge modeling and extract communication and integration into OntoCR. Blue arrows indicate knowledge-related processes, while red arrows indicate data-related processes. API: Application Programming Interface; EHR: electronic health record; ISO: International Organization for Standardization; SNOMED CT: Systematized Nomenclature of Medicine – Clinical Terms.



Results

EN/ISO 13606 EHR Extracts

We achieved effective communication of EN/ISO 13606–standardized EHR extracts between a mobile app for patients, Xemio Research, and an institutional clinical repository, OntoCR.

In our study pilot, 62 patients were allocated to use Xemio Research. There were 12 (19.4%) dropouts: 7 (58%) due to a lack of response to questionnaires, 2 (17%) due to medical issues, 2 (17%) lost to follow-up, and 1 (8%) for personal reasons. Furthermore, 3 (4.8%) patients never used the app, leading to a total of 47 (75.8%) users.

Table 1 shows the number of each type of extracts exchanged between December 22, 2020, and April 4, 2022, and the number of patients they pertain to.

When comparing the extracts to the data registered in both Xemio Research and OntoCR databases, no missing or unclear data were detected in the process for the study cohort.

Table 1. Number of extracts communicated throughout the study.

EHR[a] archetype	Extracts (N=1100), n (%)	Patients (N=47), n (%)
Side effects	234 (21.3)	34 (72.3)
Daily activity	866 (78.7)	38 (80.9)

[a]EHR: electronic health record.

Archetypes and Ontologies

Furthermore, the methodology created for this project resulted in a series of deliverables within each step of the process. First, the creation of EN/ISO 13606–standardized archetypes allows the reuse of clinical information for the variables considered in this study: daily activity (date, steps, calories, and duration) and side effects (date, finding, value, and severity).

In addition, by creating ontologies that represent the aforementioned clinical variables and integrating them into OntoCR, we continue to extend the knowledge representation of our ontology-based clinical repository.

Discussion

Principal Findings

We describe a methodology for communicating EN/ISO 13606 EHR extracts between a patient mobile app and an ontology-based clinical repository. Standardized information regarding side effects or daily activity of patients enrolled into Xemio Research in the study was effectively communicated.

EN/ISO 13606 was chosen for this project because of the operational mechanisms it offers for data exchange and its advantages regarding modeling of clinical knowledge and flexibility in the creation of new concepts, which is also why it was used in the first place to extend OntoCR's metamodel with the incorporation of the reference and archetype models of the standard. However, due to the flexibility and standard-agnostic nature of our methodology, there is complete independence regarding any specific standard. Thus, we are able to carry out transformations between health information standards with minimum use of resources and without the need for changes in the database structure.

LinkEHR offers the possibility to create clinical information models using multiple health information standards (EN/ISO 13606, OpenEHR, FHIR) as well as terminologies and classifications (SNOMED CT, International Classification of Diseases 10th Revision [ICD-10], Logical Observation Identifiers Names and Codes [LOINC]), all of which can also be incorporated into OntoCR by creating corresponding metamodel ontologies. The API that inserts instantiated patient data into the repository is prepared to receive any EN/ISO 13606 EHR extract, and it can be extended to incorporate other standards as well. All this facilitates the application of our methodology to other projects and institutions.

Single vs Dual Models and Semantic Interoperability in Health Care

Health information interoperability is one of the requirements for the continuity of health care [21]. The dual model allows the separation of knowledge and information in EHR systems, with the consequent possibility of extending the concept model without the need for specific developments and introducing new concepts when the system is already implemented [22]. With the use of formal information models built from common components and linked to standard terminologies [23], 2 systems can achieve semantic interoperability without prior agreement [24,25].

Single, nonstandardized models require the development of specific interfaces to communicate information with other systems. In a context where there is a growing number of information systems within each health organization, many of which come from mobile devices of both patients and physicians, the scalability of this approach is considerably reduced. These difficulties are even greater when considering the communication of health information between different organizations.

The benefits of standardizing EHR data are not limited to primary use. The reuse of clinical data for secondary purposes, such as investigation in both single- and multicenter studies, requires formal information models in order to make data unequivocally understandable and reproducible [26].

Comparison With Prior Work

There are reports in the literature of standard-agnostic approaches similar to ours, which enable a semantically interoperable clinical data landscape. Gaudet-Blavignac et al

[27] propose a 3-pillar strategy based on a multidimensional encoding of concepts, a resource description framework (RDF)–based storage and transport of the instances of these concepts, and a conversion of the RDF to any target data model. Likewise, the INFOBANCO project of the Madrid Region in Spain [28] aims to create a platform for the management, persistence, exchange, and reuse of health data, contemplating 2 types of outputs: interoperability (HL7 FHIR, EN/ISO 13606, Clinical Data Interchange Standards Consortium [CDISC] [29]) and persistence (OpenEHR, i2b2, Observational Medical Outcomes Partnership Common Data Model [OMOP CDM]). It uses a standard-agnostic design that seeks to apply each health information standard for the purpose it was intended to, offering multiple interoperability and exploitation services suited for specific use cases [12]. However, these projects focus on the creation of interoperable platforms for different purposes, but they do not include a strategy for integrating information coming from mobile apps.

Other groups have reported the use of the SMART on FHIR framework to integrate PGHD from mobile apps into EHRs [30-33]. This framework enables medical apps to be written once and run unmodified across different health care IT systems and has proven to be an effective approach for interoperability. FHIR offers operational mechanisms for data exchange, but unlike EN/ISO 13606, it lacks the capacity to build new concepts based on specific requirements [12], which limits its flexibility to adapt to new scenarios.

Strengths and Limitations

There are strengths to this study that are worth mentioning. First, the 3 main software programs used (LinkEHR, Protégé, and Liferay) are open source, which makes our methodology accessible to low-income areas as well as institutions with limited funding for such projects. Moreover, the aforementioned flexibility and standard-agnostic nature of our methodology define a considerable scalability. The knowledge representation can be adjusted to different contexts with little resources, just by creating new archetypes, modeling the clinical concepts, and mapping them to the corresponding structure of EN/ISO 13606. If a different health information standard is to be used, its metamodel must be represented with ontologies, and both the conversion module and the API need to be adjusted.

With a few exceptions, such as the experience reported by Zenteno et al [34], there is limited evidence in the literature regarding the effective communication and integration of EN/ISO 13606–standardized extracts from a mobile app into an EHR. In addition, to the best of our knowledge, ours is the first experience that does so with data coming from a patient mobile app. Given EN/ISO 13606's advantages over FHIR in terms of modeling and formalization of clinical knowledge and flexibility in the creation of new concepts, our approach proves to be quite innovative in the communication of EHR extracts with mobile apps.

This study also has some limitations. First, even though there is a log file in the server that registers the extracts that are sent, there is no alarm that notifies us when the process is not working. Therefore, this maintenance and update of the system still depends on manual processes. Furthermore, the ontology-based approach requires trained staff and an initial development that involves the allocation of resources in terms of personnel, funds, and time, which can limit the extensibility of the methodology to other contexts.

Next Steps

Regarding next steps of the project, we are in the process of integrating a dashboard into OntoCR, which will display the AI-based predicted variation in the QoL issues according to the interventions carried out by physicians. This will help physicians with their clinical decision-making when evaluating treatment alternatives for breast cancer patients.

Furthermore, we are working on extending the integration of extracts to other functionalities in Xemio Research, and later, we plan to do so with other mobile apps used within the HCB ecosystem.

Conclusion

This study describes a novel methodology for the successful communication of standardized EHR extracts from a patient mobile app with an ontology-based clinical repository linked to an EHR. Its flexibility and standard-agnostic nature provide significant scalability to adapt to different contexts, situations, and information systems, while the use of open source software facilitates its transferability to other institutions. Our approach allows the integration of data coming from different sources into HISs for them to be used in the caregiving process (primary use) or for investigation purposes (secondary use). To the best of our knowledge, this is the first study to achieve effective communication and integration of EN/ISO 13606–standardized extracts from a patient mobile app into an EHR.

Acknowledgments

This research work was carried out as part of the EU-funded Research and Innovation Action, Artificial intelligence Supporting CAncer Patients across Europe (ASCAPE; Project ID: 875351), [H2020-SC1-DTH-2019] SC1-DTH-01-2019, Big data and Artificial Intelligence for monitoring health status and quality of life after the cancer treatment.

The improvement of the Xemio Research platform was supported by a La Caixa Foundation grant (LCF/PR/AR19/51450002).

Conflicts of Interest

None declared.

References

1. Treadwell JR, Rouse B, Reston J, Fontanarosa J, Patel N, Mull NK. Consumer Devices for Patient-Generated Health Data Using Blood Pressure Monitors for Managing Hypertension: Systematic Review. JMIR Mhealth Uhealth 2022 May 02;10(5):e33261 [FREE Full text] [doi: 10.2196/33261] [Medline: 35499862]
2. Benson T. Why PROMs and PREMs matter? In: Patient-Reported Outcomes and Experience. Cham: Springer; Apr 30, 2022:3-12.
3. Porter ME. What is value in health care? N Engl J Med 2010 Dec 23;363(26):2477-2481. [doi: 10.1056/NEJMp1011024] [Medline: 21142528]
4. Gray M. Value based healthcare. BMJ 2017 Jan 27;356:j437. [doi: 10.1136/bmj.j437] [Medline: 28130219]
5. Meadows KA. Patient-reported outcome measures: an overview. Br J Community Nurs 2011 Mar;16(3):146-151. [doi: 10.12968/bjcn.2011.16.3.146] [Medline: 21378658]
6. Kalra D, Beale T, Heard S. The openEHR Foundation. Stud Health Technol Inform 2005;115:153-173. [Medline: 16160223]
7. International Organization for Standardization. ISO 13606 Standard, Part1: Reference Model. URL: https://www.iso.org/standard/67868.html [accessed 2022-05-13]
8. International Organization for Standardization. ISO 13606 Standard, Part 2: Archetype Interchange Specification. URL: https://www.iso.org/standard/62305.html [accessed 2022-05-13]
9. Martínez-Costa C, Menárguez-Tortosa M, Fernández-Breis JT. An approach for the semantic interoperability of ISO EN 13606 and OpenEHR archetypes. J Biomed Inform 2010 Oct;43(5):736-746 [FREE Full text] [doi: 10.1016/j.jbi.2010.05.013] [Medline: 20561912]
10. HL7. HL7 FHIR. URL: https://hl7.org/FHIR/ [accessed 2022-05-13]
11. Mandel JC, Kreda DA, Mandl KD, Kohane IS, Ramoni RB. SMART on FHIR: a standards-based, interoperable apps platform for electronic health records. J Am Med Inform Assoc 2016 Sep;23(5):899-908 [FREE Full text] [doi: 10.1093/jamia/ocv189] [Medline: 26911829]
12. Pedrera-Jiménez M, Spanish Expert Group on EHR standards, Kalra D, Beale T, Muñoz-Carrero A, Serrano-Balazote P. Can OpenEHR, ISO 13606 and HL7 FHIR work together? An agnostic perspective for the selection and application of EHR standards from Spain Internet. TechRxiv 2022 May 25 [FREE Full text] [doi: 10.36227/techrxiv.19746484]
13. Tzelves L, Manolitsis I, Varkarakis I, Ivanovic M, Kokkonidis M, Useros C, et al. Artificial intelligence supporting cancer patients across Europe-The ASCAPE project. PLoS One 2022;17(4):e0265127 [FREE Full text] [doi: 10.1371/journal.pone.0265127] [Medline: 35446854]
14. Savić M, Kurbalija V, Ilić M, Ivanović M, Jakovetić D, Valachis A. Analysis of machine learning models predicting quality of life for cancer patients. 2021 Nov 09 Presented at: MEDES'21: 13th International Conference on Management of Digital EcoSystems; November 1-3, 2021; Virtual p. 35-42. [doi: 10.1145/3444757.3485103]
15. Lampropoulos K, Kosmidis T, Autexier S, Savic M, Athanatos M, Kokkonidis M. ASCAPE: an open AI ecosystem to support the quality of life of cancer patients Internet. 2021 Oct 15 Presented at: 2021 IEEE 9th International Conference on Healthcare Informatics (ICHI); August 9-12, 2021; Victoria, BC, Canada. [doi: 10.1109/ichi52183.2021.00054]
16. Lozano-Rubí R, Muñoz Carrero A, Serrano Balazote P, Pastor X. OntoCR: A CEN/ISO-13606 clinical repository based on ontologies. J Biomed Inform 2016 Apr;60:224-233 [FREE Full text] [doi: 10.1016/j.jbi.2016.02.007] [Medline: 26911524]
17. Lozano-Rubí R. A Metamodel for Clinical Data Integration: Basis for a New EHR Model Driven by Ontologies. 2016 Nov 11. URL: https://www.tdx.cat/bitstream/handle/10803/399855/rlr1de1.pdf?sequence=1 [accessed 2022-09-30]
18. Sezov R. Liferay in Action: The Official Guide to Liferay Portal Development. New York, NY: Simon and Schuster; 2011.
19. Nocentino A, Weissman B. Kubernetes architecture. In: SQL Server on Kubernetes. Berkeley, CA: Apress; 2021.
20. Maldonado JA, Moner D, Boscá D, Fernández-Breis JT, Angulo C, Robles M. LinkEHR-Ed: a multi-reference model archetype editor based on formal semantics. Int J Med Inform 2009 Aug;78(8):559-570. [doi: 10.1016/j.ijmedinf.2009.03.006] [Medline: 19386540]
21. Muñoz A, Somolinos R, Pascual M, Fragua JA, González MA, Monteagudo JL, et al. Proof-of-concept design and development of an EN13606-based electronic health care record service. J Am Med Inform Assoc 2007;14(1):118-129 [FREE Full text] [doi: 10.1197/jamia.M2058] [Medline: 17068357]
22. Blobel B. Advanced and secure architectural EHR approaches. Int J Med Inform 2006;75(3-4):185-190. [doi: 10.1016/j.ijmedinf.2005.07.017] [Medline: 16112891]
23. Goossen WTF. Detailed clinical models: representing knowledge, data and semantics in healthcare information technology. Healthc Inform Res 2014 Jul;20(3):163-172 [FREE Full text] [doi: 10.4258/hir.2014.20.3.163] [Medline: 25152829]
24. Moreno-Conde A, Moner D, Cruz WDD, Santos MR, Maldonado JA, Robles M, et al. Clinical information modeling processes for semantic interoperability of electronic health records: systematic review and inductive analysis. J Am Med Inform Assoc 2015 Jul;22(4):925-934. [doi: 10.1093/jamia/ocv008] [Medline: 25796595]
25. Pedrera-Jiménez M, García-Barrio N, Cruz-Rojo J, Terriza-Torres AI, López-Jiménez EA, Calvo-Boyero F, et al. Obtaining EHR-derived datasets for COVID-19 research within a short time: a flexible methodology based on Detailed Clinical Models. J Biomed Inform 2021 Mar;115:103697 [FREE Full text] [doi: 10.1016/j.jbi.2021.103697] [Medline: 33548541]
26. Pedrera M, Garcia N, Rubio P, Cruz J, Bernal J, Serrano P. Making EHRs Reusable: A Common Framework of Data Operations. Stud Health Technol Inform 2021 Nov 18;287:129-133. [doi: 10.3233/SHTI210831] [Medline: 34795096]

27. Gaudet-Blavignac C, Raisaro JL, Touré V, Österle S, Crameri K, Lovis C. A National, Semantic-Driven, Three-Pillar Strategy to Enable Health Data Secondary Usage Interoperability for Research Within the Swiss Personalized Health Network: Methodological Study. JMIR Med Inform 2021 Jun 24;9(6):e27591 [FREE Full text] [doi: 10.2196/27591] [Medline: 34185008]
28. Comunidad de Madrid. Infobanco. URL: https://cpisanidadcm.org/infobanco/ [accessed 2022-09-30]
29. Clinical Data Interchange Standards Consortium. CDISC Standards in the Clinical Research Process. URL: https://www.cdisc.org/standards [accessed 2022-09-30]
30. Spratt S, Ravneberg D, Derstine B, Granger B. Feasibility of Electronic Health Record Integration of a SMART Application to Facilitate Patient-Provider Communication for Medication Management. Comput Inform Nurs 2022 Aug 01;40(8):538-546. [doi: 10.1097/CIN.0000000000000891] [Medline: 35234708]
31. Curran R, Kukhareva P, Taft T, Weir C, Reese T, Nanjo C, et al. Integrated displays to improve chronic disease management in ambulatory care: A SMART on FHIR application informed by mixed-methods user testing. J Am Med Inform Assoc 2020 Aug 01;27(8):1225-1234 [FREE Full text] [doi: 10.1093/jamia/ocaa099] [Medline: 32719880]
32. Sayeed R, Gottlieb D, Mandl KD. SMART Markers: collecting patient-generated health data as a standardized property of health information technology. NPJ Digit Med 2020;3:9 [FREE Full text] [doi: 10.1038/s41746-020-0218-6] [Medline: 31993507]
33. Wesley D, Blumenthal J, Shah S, Littlejohn R, Pruitt Z, Dixit R, et al. A novel application of SMART on FHIR architecture for interoperable and scalable integration of patient-reported outcome data with electronic health records. J Am Med Inform Assoc 2021 Sep 18;28(10):2220-2225 [FREE Full text] [doi: 10.1093/jamia/ocab110] [Medline: 34279660]
34. Torres Zenteno AH, Fernández F, Palomino-García A, Moniche F, Escudero I, Jiménez-Hernández MD, et al. Mobile platform for treatment of stroke: A case study of tele-assistance. Health Informatics J 2016 Sep;22(3):676-690 [FREE Full text] [doi: 10.1177/1460458215572925] [Medline: 25975806]

Abbreviations

AI: artificial intelligence
API: Application Programming Interface
ASCAPE: Artificial intelligence Supporting CAncer Patients across Europe
CDISC: Clinical Data Interchange Standards Consortium
EHR: electronic health record
EN/ISO: European Norm/International Organization for Standardization
FHIR: Fast Healthcare Interoperability Resources
FL: federated learning
GUI: graphical user interface
HCB: Hospital Clínic de Barcelona
HIS: health information system
HL7: Health Level 7
ICD-10: International Classification of Diseases 10th Revision
IT: information technology
LOINC: Logical Observation Identifiers Names and Codes
ML: machine learning
OMOP CDM: Observational Medical Outcomes Partnership Common Data Model
PGHD: patient-generated health data
PROM: patient-reported outcome measure
QoL: quality of life
SMART: Substitutable Medical Applications and Reusable Technologies
SNOMED CT: Systematized Nomenclature of Medicine – Clinical Terms
URI: uniform resource identifier

Edited by C Lovis, J Hefner; submitted 16.06.22; peer-reviewed by C Gaudet-Blavignac, O Endrich, M Pedrera Jiménez, A Muñoz, C Rodríguez; comments to author 26.07.22; revised version received 12.08.22; accepted 01.09.22; published 12.10.22.

Please cite as:
Frid S, Fuentes Expósito MA, Grau-Corral I, Amat-Fernandez C, Muñoz Mateu M, Pastor Duran X, Lozano-Rubí R
Successful Integration of EN/ISO 13606–Standardized Extracts From a Patient Mobile App Into an Electronic Health Record: Description of a Methodology
JMIR Med Inform 2022;10(10):e40344
URL: https://medinform.jmir.org/2022/10/e40344
doi:10.2196/40344
PMID:36222792

Original Paper

Fast Healthcare Interoperability Resources for Inpatient Deterioration Detection With Time-Series Vital Signs: Design and Implementation Study

Tzu-Wei Tseng[1*], MSc; Chang-Fu Su[2*], MD; Feipei Lai[1*], PhD

[1]Department of Computer Science and Information Engineering, National Taiwan University, Taipei City, Taiwan

[2]Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei City, Taiwan

[*]all authors contributed equally

Corresponding Author:
Tzu-Wei Tseng, MSc
Department of Computer Science and Information Engineering
National Taiwan University
No. 1, Sec. 4, Roosevelt Rd.
Taipei City, 106319
Taiwan
Phone: 886 963079621
Email: chaaa463@gmail.com

Abstract

Background: Vital signs have been widely adopted in in-hospital cardiac arrest (IHCA) assessment, which plays an important role in inpatient deterioration detection. As the number of early warning systems and artificial intelligence applications increases, health care information exchange and interoperability are becoming more complex and difficult. Although Health Level 7 Fast Healthcare Interoperability Resources (FHIR) have already developed a vital signs profile, it is not sufficient to support IHCA applications or machine learning–based models.

Objective: In this paper, for IHCA instances with vital signs, we define a new implementation guide that includes data mapping, a system architecture, a workflow, and FHIR applications.

Methods: We interviewed 10 experts regarding health care system integration and defined an implementation guide. We then developed the FHIR Extract Transform Load to map data to FHIR resources. We also integrated an early warning system and machine learning pipeline.

Results: The study data set includes electronic health records of adult inpatients who visited the En-Chu-Kong hospital. Medical staff regularly measured these vital signs at least 2 to 3 times per day during the day, night, and early morning. We used pseudonymization to protect patient privacy. Then, we converted the vital signs to FHIR observations in the JSON format using the FHIR Extract Transform Load application. The measured vital signs include systolic blood pressure, diastolic blood pressure, heart rate, respiratory rate, and body temperature. According to clinical requirements, we also extracted the electronic health record information to the FHIR server. Finally, we integrated an early warning system and machine learning pipeline using the FHIR RESTful application programming interface.

Conclusions: We successfully demonstrated a process that standardizes health care information for inpatient deterioration detection using vital signs. Based on the FHIR definition, we also provided an implementation guide that includes data mapping, an integration process, and IHCA assessment using vital signs. We also proposed a clarifying system architecture and possible workflows. Based on FHIR, we integrated the 3 different systems in 1 dashboard system, which can effectively solve the complexity of the system in the medical staff workflow.

(JMIR Med Inform 2022;10(10):e42429) doi:10.2196/42429

KEYWORDS

Fast Healthcare Interoperability Resources; FHIR; Health Level 7; HL7; health research; data sharing; health information technology; clinical research

Introduction

Background

Vital signs have been an important indicator in many studies [1-3]. In recent years, researchers have used these data in studies of predictive models for in-hospital cardiac arrest (IHCA) [1,4]. In a real-world medical workflow, complete data may be obtained once every 4 to 8 hours. In the machine learning research related to vital signs [5], the features include heart rate, temperature, respiratory rate, systolic blood pressure, and diastolic blood pressure. In addition to IHCA risk assessment, data analysis systems [6] and early warning systems [7] are still indispensable applications.

Although IHCA risk indicators have facilitated breakthroughs in machine learning [8,9], it has been difficult to integrate them into the workflow of medical staff. In hospitals, there are dozens of systems that must exchange information with each other. Without a standard exchange interface [10], the integration process is costly and time-consuming when a new application is imported. In addition, if medical researchers are allowed to access patient data directly through the health care information system database, security risks [11] become a concern.

To begin initiating a human-readable and user-friendly interface for medical staff, Health Level 7 [12] developed Fast Healthcare Interoperability Resources (FHIR) [13]. FHIR is a platform specification that defines a set of capabilities used across the health care processes, and it defines a generic health care business entity model that uses resources as the basic blocks. Each resource in FHIR has a defined relationship resource with data elements and constraints. In addition, the FHIR profile standardizes the data format and structure constraints. During data transportation, it uses the HTTP RESTful application programming interface (API) in the exchange interface and provides the flexibility to choose between JSON or XML format in the data payload.

Aim

Although FHIR have developed some of the resources, a vital signs profile [14] has not yet matured. The current implementation guide provided by FHIR is insufficient to encompass the full range of medical system applications; therefore, hospitals still need to define the customized implementation guide to develop their system and workflow. The implementation guide is a collection of rules applied by FHIR resources [15] that requires a clear explanation of how to solve a particular problem. In the relevant studies on FHIR [16-18], each paper develops and discusses a single customized resource profile on a mobile device. Although FHIR can effectively and rapidly improve health care information system interoperability, it still has not proposed an implementation guide for the machine learning application in FHIR implementation guide registry. To accelerate the development of smart health care, we propose a system architecture process based on FHIR that can integrate the machine learning models. Besides, the vital signs applications are distributed in many different systems. This study can effectively solve the complexity of the system in the medical staff workflow.

To standardize the format among medical systems, we developed a complete IHCA implementation guide based on FHIR that defines the vital signs–related data for both the early warning system and the machine learning pipeline. In addition, we also developed FHIR Extract Transform Load (ETL) and other FHIR-related applications, including data management, an early warning system, and a machine learning pipeline.

Methods

Ethics Approval

This study was approved by the Institutional Review Board of the En-Chu-Kong Hospital (ECKIRB1071001). We confirm that all experiments were performed in accordance with relevant guidelines and regulations. The data retrieved from electronic health records (EHRs) were deidentified by an IT specialist and could not be linked to the patients' identity by the research team. The need for written informed consent was waived and confirmed by the En-Chu-Kong Hospital Institutional Review Board, because this was a retrospective cohort study with deidentified data.

Overview

Our study provides a design and implementation process for IHCA-based interoperability of health care information systems, and our design steps include use cases as well as the IHCA implementation guide.

In the use cases section, we describe the integration issues faced by health care institutions. Then, in the IHCA implementation guide section, we introduce the method used to migrate data from the healthcare information system (HIS) database to the FHIR server as well as a method for mapping the data to the FHIR resources. We also develop the 3 application systems, which include data management, early warning systems, and a machine learning pipeline. If used according to our implementation guide, the applications can easily obtain patient information and vital signs data.

Use Case Survey

In health care institutions, the database is centrally managed, but the applications are developed by many different teams. In addition, medical staff usually access all of the required information about a workflow through a single system. Therefore, the interoperability of health care systems is very important.

To achieve system information interoperability [19], the HTTP RESTful API was defined to exchange data with other systems. However, many medical systems are legacy systems, and in many cases, it is impossible to change the system architecture. We therefore created a table view for the HIS database to allow other systems to obtain particular data. To avoid affecting the original system architecture, we developed FHIR ETL to convert data from the HIS database to the FHIR server, and FHIR ETL was implemented according to the rules defined by the IHCA implementation guide.

We interviewed 10 experts regarding health care system integration and information exchange. As shown in Table 1, FHIR, which has a good medical standard interface, is very

suitable for solving the interoperability problems faced by medical information systems. In addition, it supports a variety of systems that can be used to develop extended applications.

Therefore, we have 2 use cases. The first use case is related to data migration for the FHIR server, as shown in Figure 1 (Part A). The second use case is related to FHIR applications, as shown in Figure 1 (Part B).

Table 1. Requirement list from health care specialists in health care institutions.

Issue (requirement)	How to do it?
The new system integration process shall not affect the health care information system or the vital signs system.	Build the FHIR[a] server as a new middleware or gateway so that researchers can access data.
Converting the EHRs[b] with vital signs into FHIR resources.	Develop the FHIR ETL[c].
To reduce the time cost and compatibility, we need to use a health care information interoperability standard.	Use FHIR resources and the RESTful API[d].
The field needs an early warning system that can continuously monitor the patient's vital signs.	Use FHIR to develop the early warning system.
How can an organization integrate the prediction model into the medical workflow?	Use FHIR to develop the machine learning pipeline.
The field needs a complete implementation procedure and use case.	Define an FHIR implementation guide.

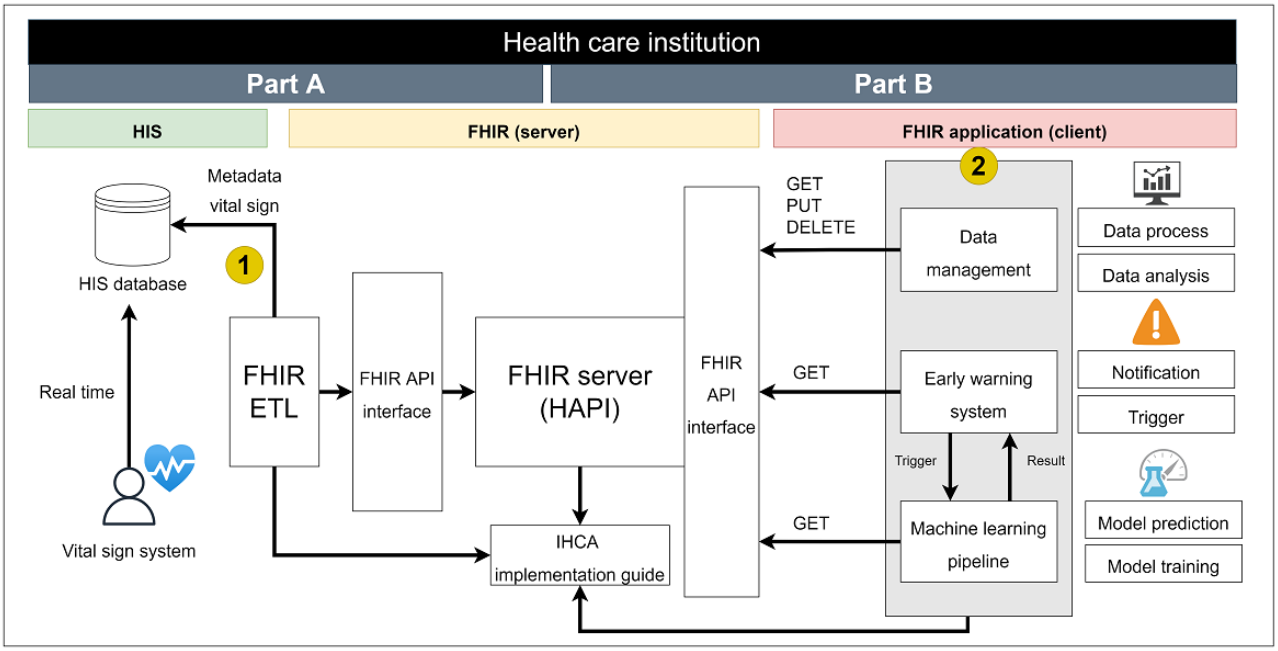
[a]FHIR: Fast Healthcare Interoperability Resources.

[b]EHRs: electronic health records.

[c]ETL: Extract Transform Load.

[d]API: application programming interface.

Figure 1. Use cases for IHCA research and application. (A) Extract the data and transfer them to the FHIR server. (B) Data management for data processing, early warning system for notification and model trigger, and machine learning pipeline for model prediction and model training. API: application programming interface; ETL: Extract Transform Load; FHIR: Fast Healthcare Interoperability Resources; HAPI: Health Level 7 application programming interface; HIS: healthcare information system; IHCA: in-hospital cardiac arrest.



IHCA Implementation Guide

In this phase, we need to consider the data format so that raw data can be transferred into FHIR resources as well as how the HTTP RESTful API can be used to easily obtain data. Therefore, we designed a system architecture (Figure 1). We divided the system steps into the following: (1) the FHIR ETL performs data conversion and comparisons between the HIS database and the FHIR server, and (2) the application system accesses data directly through the FHIR API interface at the HTTP layer.

Data Mapping—FHIR ETL

We proposed the data mapping table to develop the FHIR ETL, as shown in Table 2. We defined the data mapping and resource relations. Based on the FHIR vital signs profile, we used the observation resource to store systolic blood pressure, diastolic blood pressure, heart rate, respiratory rate, and body temperature. According to Table 2, FHIR ETL can extract the data from the HIS database and transfer them to resource content.

Table 2. The data mapping table of the FHIR[a] ETL[b] in this study.

Data model of HIS[c] database	FHIR resource name	FHIR resource attribute	Description
Patient_ID	Patient	identifier	An identifier for the patient in the hospital
Patient_name	Patient	name	Patient's name that is human-readable
Gender	Patient	gender	Patient's gender
BirthDate	Patient	birthDate	Patient's birth date
Practitioner_ID	Practitioner	identifier	An identifier for the physician in the hospital
Practitioner_name	Practitioner	name	Physician's name that is human-readable
Organization_ID	Organization	identifier	An identifier for the department in the hospital
Organization_name	Organization	name	Department's name that is human-readable
Location_ID	Location	identifier	An identifier for the location in the hospital
Location_name	Location	name	Location's name that is human-readable
Heart rate	Observation	valueQuantity.value	Heart rate
Temperature	Observation	valueQuantity.value	Temperature
Respiratory rate	Observation	valueQuantity.value	Respiratory rate
Systolic blood pressure	Observation	valueQuantity.value	Systolic blood pressure
Diastolic blood pressure	Observation	valueQuantity.value	Diastolic blood pressure
Timestamp	Observation	effectiveDateTime	The created time of the value

[a]FHIR: Fast Healthcare Interoperability Resources.

[b]ETL: Extract Transform Load.

[c]HIS: healthcare information system.

In Figure 2, in terms of data acquisition, if an FHIR client wants to obtain a patient's location, it needs to first obtain the patient's ID and join the encounter subject. Then, it can use the encounter location to find the location resource. Finally, the FHIR client can obtain the patient location.

In Figure 3, the FHIR client can perform the following: (1) when an FHIR client needs to access a particular patient using metadata, it can use the HTTP GET method to obtain the Bundle resource response; (2) when an FHIR client wants to update the location name for the hospital, it can use the HTTP PUT method to update the Location resource; and (3) after the FHIR client obtains sufficient vital signs data from the Observation resource, it can use the HTTP DELETE method to delete the resource that is missing vital signs values.

Figure 2. Data mapping and resource relationships in the IHCA implementation guide. ETL: Extract Transform Load; FHIR: Fast Healthcare Interoperability Resources; HIS: healthcare information system; IHCA: in-hospital cardiac arrest.

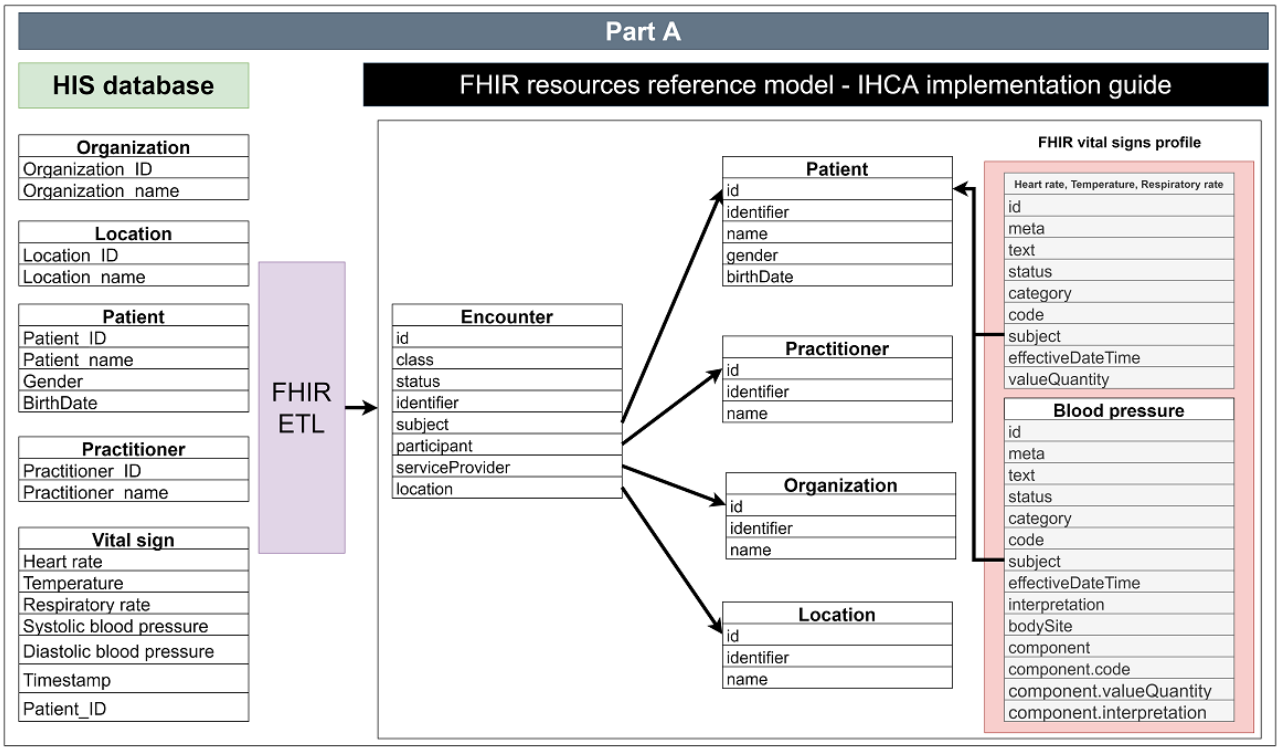
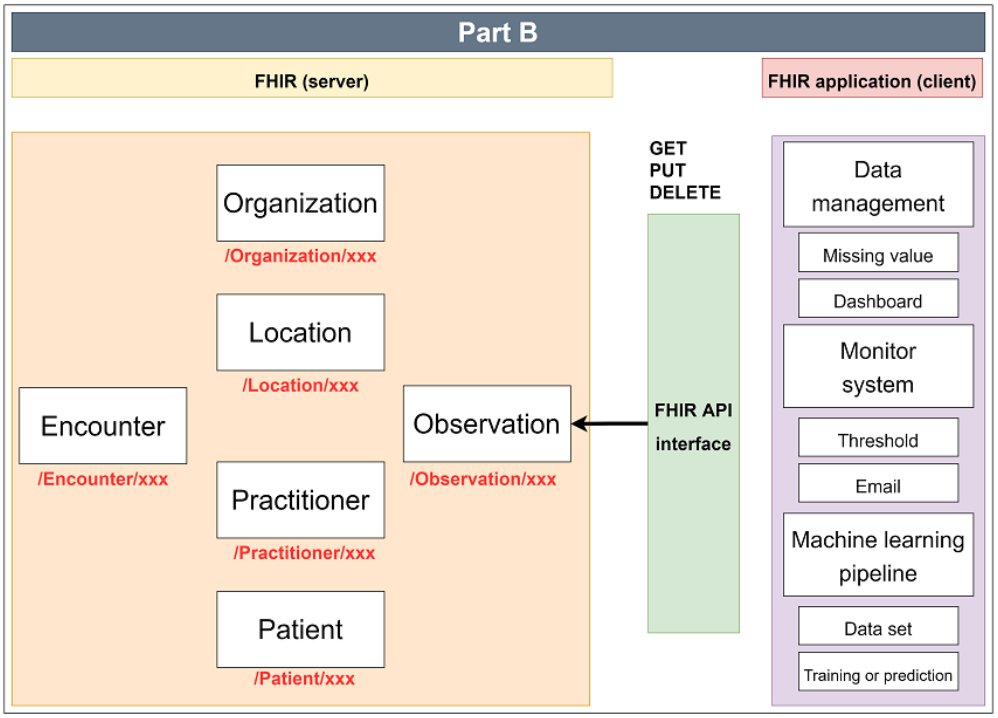


Figure 3. FHIR (Fast Healthcare Interoperability Resources) application, which uses the HTTP RESTful API (application programming interface) to control the data on the FHIR server.



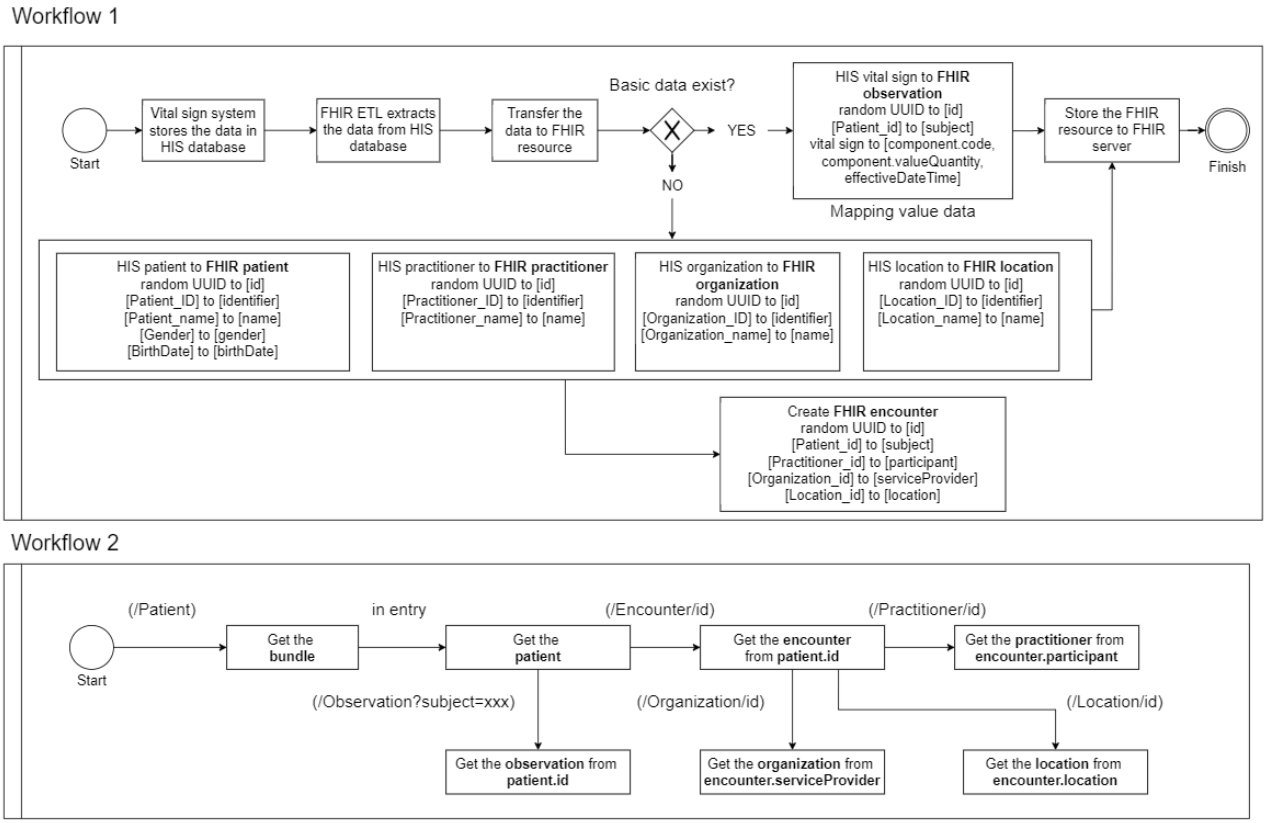
Workflow Design

In this section, we describe the complete workflow of FHIR implementation. Workflow 1 develops the data mappings for the FHIR resources. First, the FHIR ETL uses the database connection library to access the table view of the HIS database. Then, it verifies that the patient's information exists. To maintain data consistency, when converting to the Observation resource, the system must add the universally unique identifier of Patient resource as a reference link. If the patient's basic data already exists, the vital signs will be converted into an Observation resource based on the FHIR vital signs profile.

Workflow 2 develops the data acquisition process for FHIR applications. First, the FHIR application can use URL (/Patient)

with the HTTP GET method to access the Bundle resource. In the Bundle resource, the FHIR application can find all of the patient's data. If the FHIR application needs to obtain patient information such as location and practitioner information, it can use the Patient ID to join the Encounter subject. Then, it can obtain the Encounter participant and Encounter location. Finally, it can also use the Patient ID to join the URL (/Observation?subject=) with the HTTP GET method to obtain the Observation resource (Figure 4).

Figure 4. Workflow of the Fast Healthcare Interoperability Resources (FHIR) Extract Transform Load (ETL) and the FHIR client application. HIS: healthcare information system; UUID: universally unique identifier.



Results

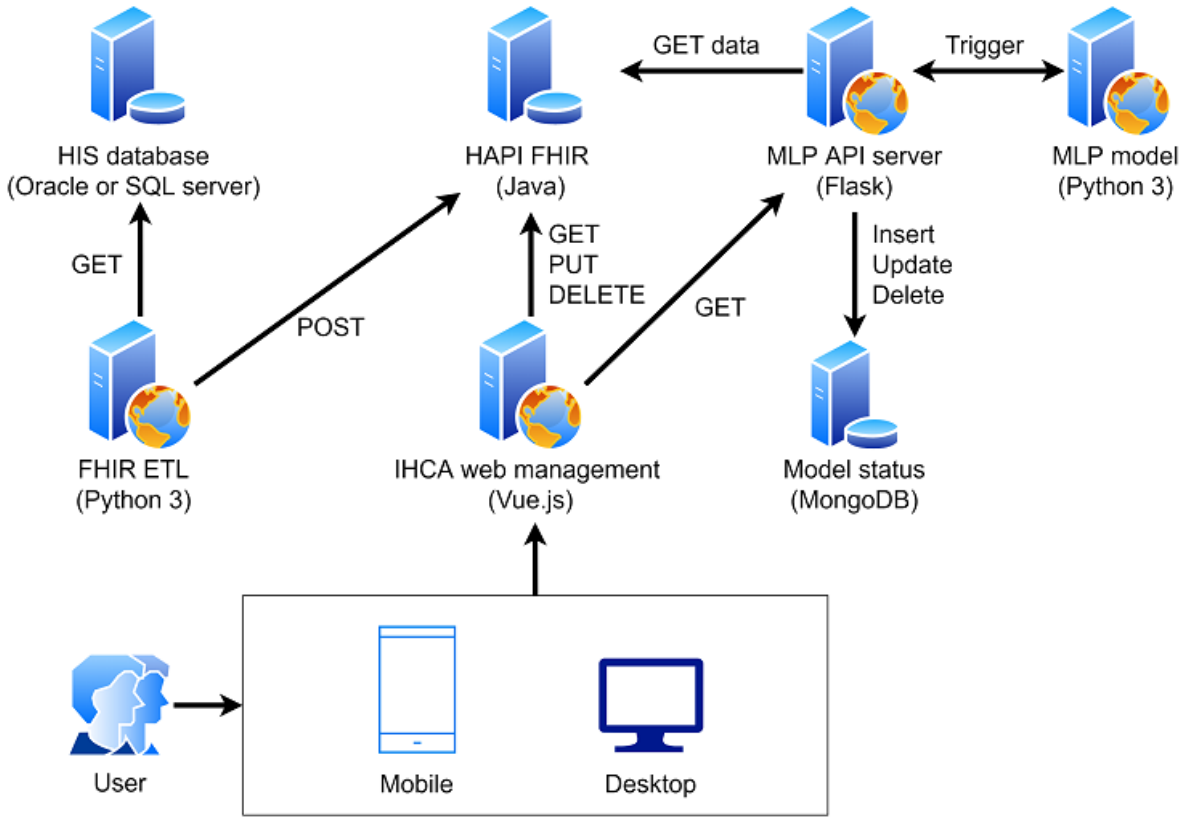
FHIR Resources

The FHIR ETL is an automation service that extracts vital signs. When the vital signs system stores the data in the HIS database, the FHIR ETL can access the vital signs data immediately, and as shown in Figure 2, it adds the vital signs to the Observation resource. Multimedia Appendix 1 shows examples of an FHIR resource that refers to an FHIR vital signs profile and other resources.

Software Development

We describe the software development, which is shown in Figure 5. The HIS database was developed using the SQL server database and the Oracle database server. The FHIR server was installed on the Health Level 7 API FHIR R4 server (version 6.1.0) [20] with a docker container based on the Java environment. This open-source system is widely used. We developed the back-end service of the FHIR ETL using Python software (version 3; Python Software Foundation), and the machine learning pipeline was implemented using Flask. The front-end website was constructed using Vue.js and is used for IHCA web management.

Figure 5. System architecture used in this study. API: application programming interface; ETL: Extract Transform Load; FHIR: Fast Healthcare Interoperability Resources; HAPI: Health Level 7 application programming interface; HIS: healthcare information system; IHCA: in-hospital cardiac arrest; MLP: machine learning pipeline.



System Implementation

The study data set [21] included the EHRs of adult inpatients who visited the En-Chu-Kong hospital. Medical staff regularly measured these vital signs at least 2 to 3 times per day during the day, night, and early morning. The total number of patients was 16,865, and the number of patients with IHCA was 118.

We converted the 5 vital signs into FHIR observations in JSON format using FHIR ETL. These vital signs include systolic blood pressure, diastolic blood pressure, heart rate, respiratory rate, and body temperature. For demonstration, we used pseudonymization [22] to protect the patient's privacy. Furthermore, we divided the proposed system into the following 3 components: data management, an early warning system, and a machine learning pipeline. In terms of data management, as shown in Figure 6, we developed a data static dashboard so that it can be accessed by medical staff using a browser. The dashboard uses the HTTP GET method to obtain both the Patient and Observation resources. Then, the patient's vital signs over the previous 48 hours are displayed. In the early warning system, medical staff can set the vital signs alert threshold to decide whether to show the alert in the notification list as shown in Figure 7. Then, the machine learning pipeline exports the vital signs data from the Observation resource to the FHIR server. We integrated a long short-term memory network–based model [21] using vital signs data to predict IHCA. It used the time series early warning score, which used heart rate, systolic blood pressure, and respiratory data. When the training process of the prediction model is initiated, the status "in progress" will appear in MongoDB. After model training, the status will be updated to "final," and the dashboard will show the latest accuracy of the model. The proposed dashboard is shown in Figure 8. However, the system can be used on mobile devices as well as desktop computers. We followed the Responsive Web Design [23] to design a user-friendly mobile interface (Figure 9).

Figure 6. Screenshot of the data management overview in the dashboard.

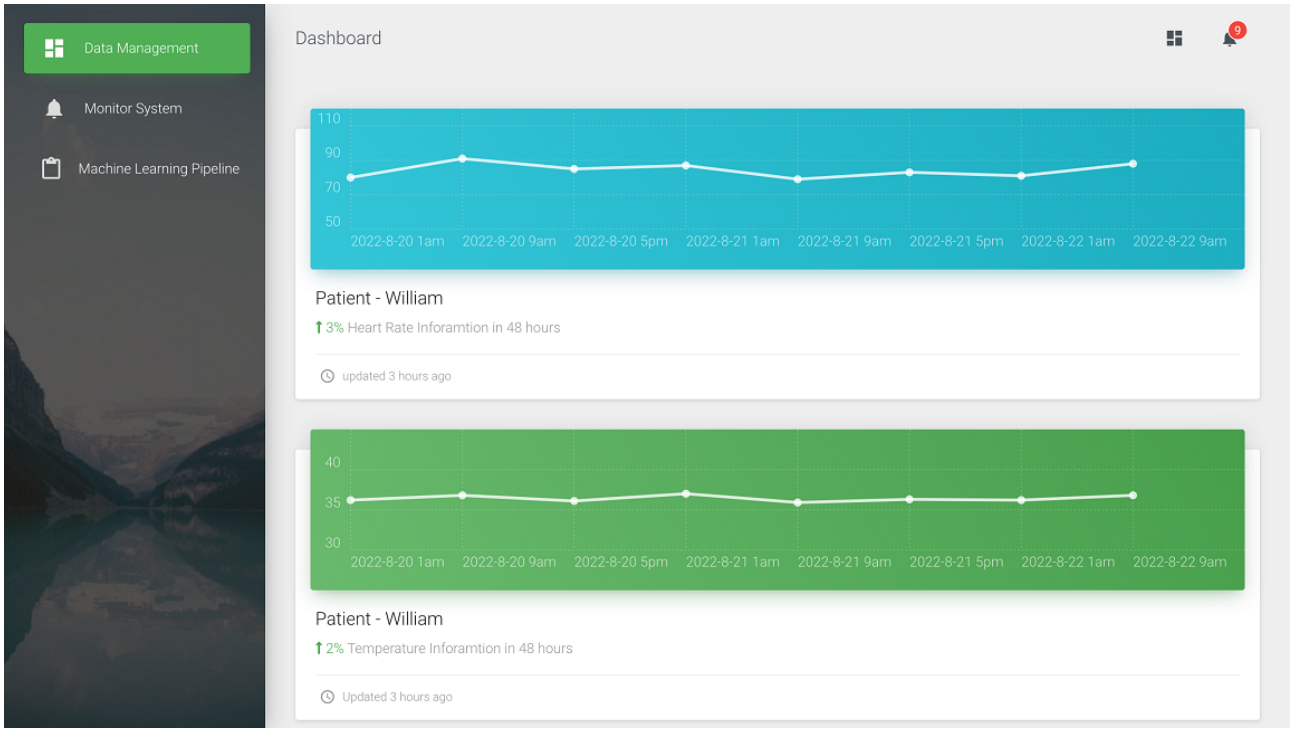


Figure 7. Screenshot of the early warning system's notification overview.

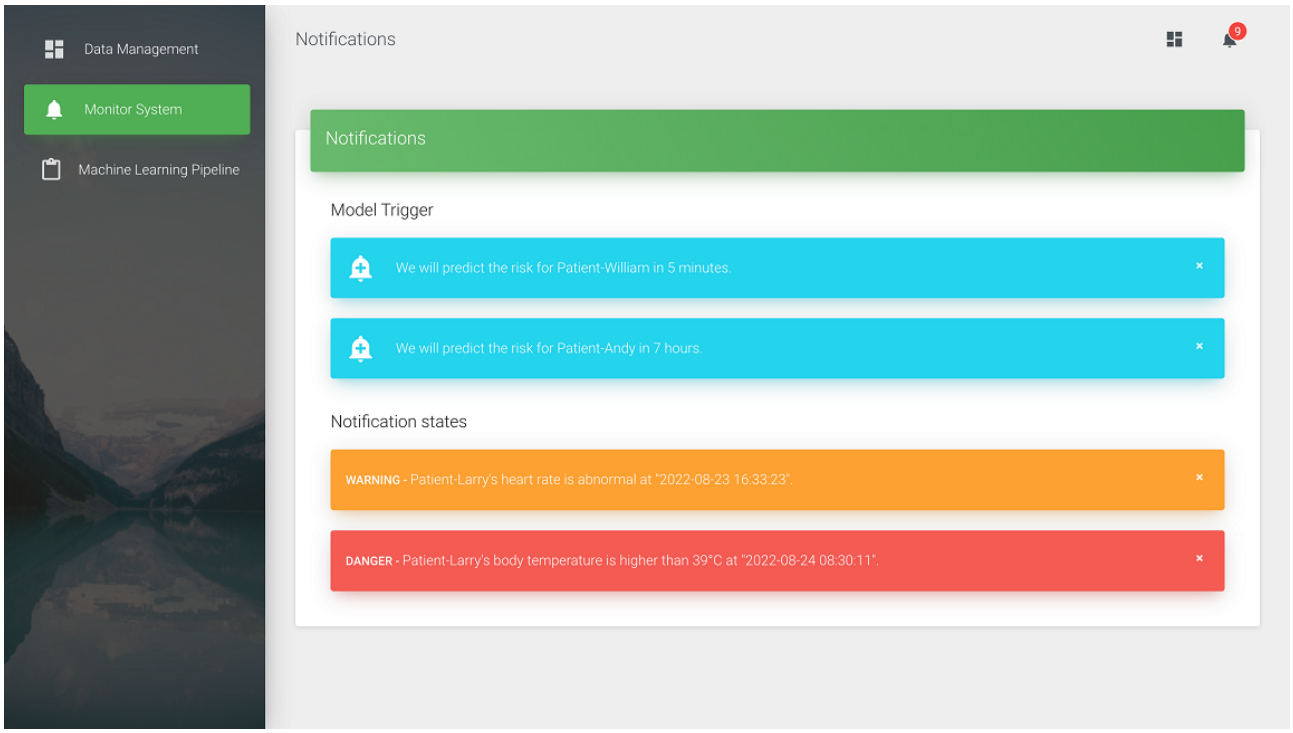


Figure 8. Screenshot of the machine learning pipeline including prediction and training. DEP: department; IHCA: in-hospital cardiac arrest.

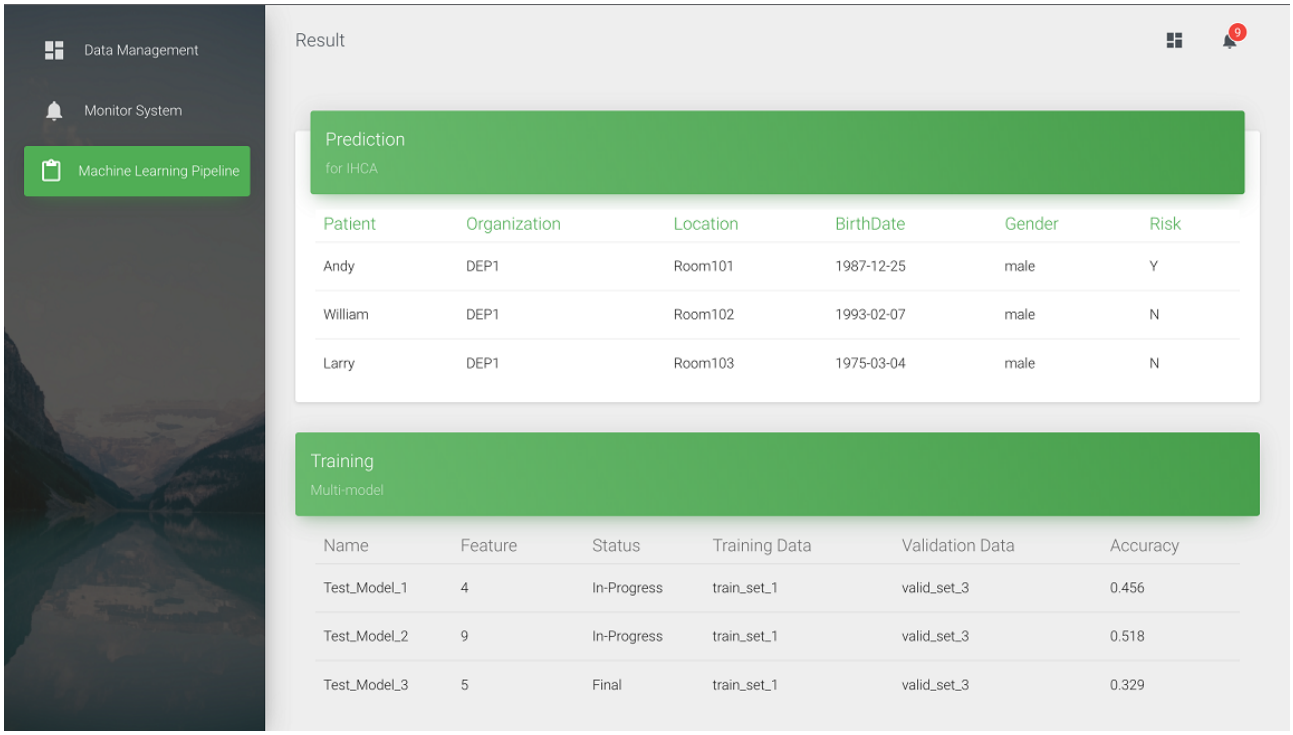
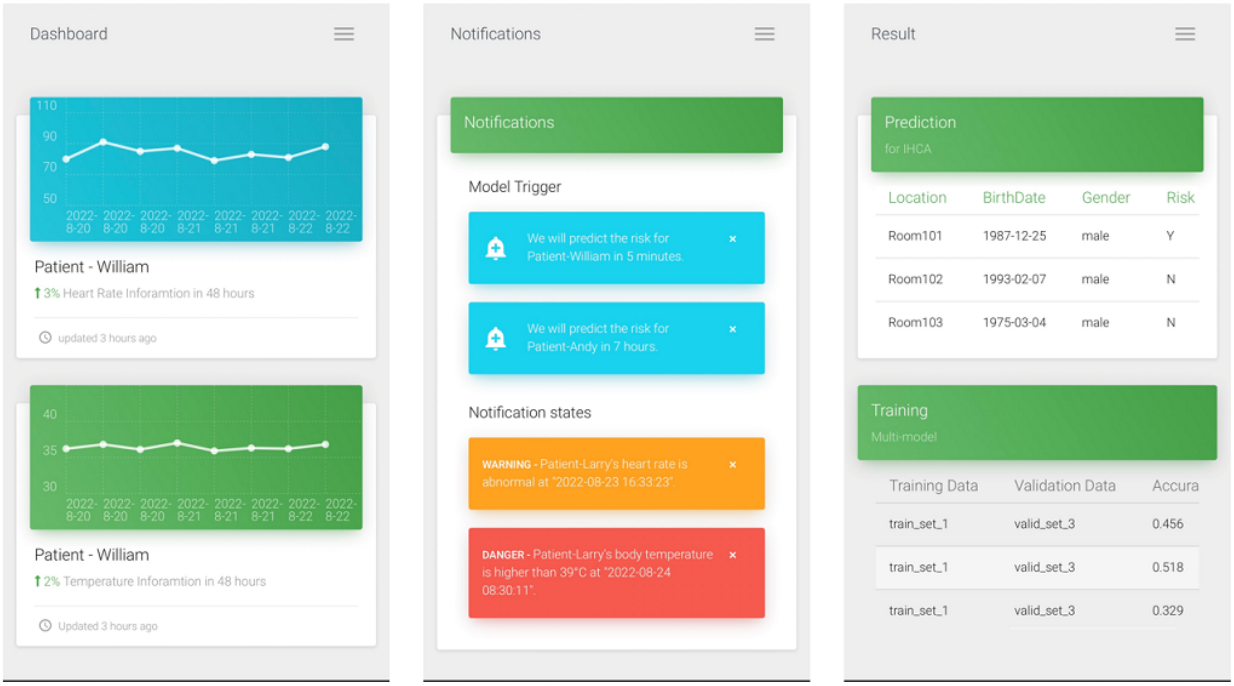


Figure 9. User interface developed using Responsive Web Design for mobile devices. IHCA: in-hospital cardiac arrest.



Discussion

Principal Findings

In this paper, we piloted the use of an implementation guide that combines IHCA with vital signs, which have been widely adopted in IHCA assessment [4,21] and play an important role in inpatient deterioration detection. Many health care institutions have developed early warning score systems to identify hospitalized patients that are at risk of deterioration, and in recent years, they have begun to incorporate machine learning–based models into this process. To promote system interoperability, we used the FHIR standard to achieve consistent information exchange. We also combined 5 resources (Organization, Location, Practitioner, Patient, and Encounter) to represent the EHR. Then, based on the FHIR vital signs profile, we exported vital signs data to HIS database and defined the IHCA implementation. In addition, we developed the 3 FHIR applications of data management using a dashboard, a real-time early warning system, and a machine learning–based pipeline. According to the IHCA implementation guide, our proposed system makes it easy to integrate vital signs–related applications.

Limitations

The implementation guide was only developed for vital signs–related studies. However, some case studies still need to include treatment history [24], blood urea nitrogen [25], and creatinine [25]. These further improvements can be made to the EHR.

Comparison With Prior Work

Despite the result that indicated that FHIR can improve the interoperability of health care information systems [26-28], existing studies have only developed the resource and profiles. Seong et al [16] demonstrated how quality information regarding clinical next-generation sequencing genomic testing can be exchanged in a standardized format by profiling an FHIR genomic resource and developing an FHIR-based web application that exchanges quality information. Based on the human-centered design methodology, Park et al [17] developed a worker-centered personal health record (PHR) app for occupational health. The PHRs were managed through a cloud server using Azure API for FHIR, and the PHR FHIR resources included Patient, Organization, DiagnosticReport, Observation, Practitioner, Condition, Procedure, MedicationStatement, Medication, and Encounter. In addition, Chukwu et al [18] profiled FHIR resources for maternal and child health referral use cases. Our study is distinguished from these previous works because we provided the IHCA implementation guidance regarding the use of FHIR resources as a conduit for the data required by the early monitoring system and machine learning. We also proposed a minimum requirements data model and combined it with the FHIR standard. To integrate the early monitoring system and machine learning, we based them on the FHIR vital sign profile and many FHIR resources to extend the data model. Besides, the related studies focus on new application development. In this study, we focus on legacy system integration, so we transfer and synchronize data through FHIR ETL.

Conclusions

We successfully demonstrated a process that standardizes health care information for inpatient deterioration detection using vital signs. Based on the FHIR definition, we provided an implementation guide that includes data mapping, an integration process, and IHCA assessment using vital signs. We also provided a clarified system architecture that can be used to develop clinical decision support systems. Based on FHIR, we integrated the 3 different systems into 1 dashboard system, which can effectively solve the complexity of the system in the medical staff workflow.

Acknowledgments

This paper was partly supported by the Ministry of Science and Technology, Taiwan (grant 10X-62634-F-002-015). The authors acknowledge the support.

Conflicts of Interest

None declared.

Multimedia Appendix 1
All Health Level 7 Fast Healthcare Interoperability Resources Bundle responses in this study.
[DOCX File , 30 KB - medinform_v10i10e42429_app1.docx]

References

1. Podell J, Pergakis M, Yang S, Felix R, Parikh G, Chen H, et al. Leveraging continuous vital sign measurements for real-time assessment of autonomic nervous system dysfunction after brain injury: a narrative review of current and future applications. Neurocrit Care 2022 Aug;37(Suppl 2):206-219. [doi: 10.1007/s12028-022-01491-6] [Medline: 35411542]
2. Yanamala N, Krishna NH, Hathaway QA, Radhakrishnan A, Sunkara S, Patel H, et al. A vital sign-based prediction algorithm for differentiating COVID-19 versus seasonal influenza in hospitalized patients. NPJ Digit Med 2021 Jun 04;4(1):95 [FREE Full text] [doi: 10.1038/s41746-021-00467-8] [Medline: 34088961]
3. Youssef Ali Amer A, Wouters F, Vranken J, de Korte-de Boer D, Smit-Fun V, Duflot P, et al. Vital signs prediction and early warning score calculation based on continuous monitoring of hospitalised patients using wearable technology. Sensors (Basel) 2020 Nov 18;20(22):6593 [FREE Full text] [doi: 10.3390/s20226593] [Medline: 33218084]
4. Chae M, Han S, Gil H, Cho N, Lee H. Prediction of in-hospital cardiac arrest using shallow and deep learning. Diagnostics (Basel) 2021 Jul 13;11(7):1255 [FREE Full text] [doi: 10.3390/diagnostics11071255] [Medline: 34359337]
5. Alghatani K, Ammar N, Rezgui A, Shaban-Nejad A. Predicting intensive care unit length of stay and mortality using patient vital signs: machine learning model development and validation. JMIR Med Inform 2021 May 05;9(5):e21347 [FREE Full text] [doi: 10.2196/21347] [Medline: 33949961]
6. Fan Y, Xu P, Jin H, Ma J, Qin L. Vital sign measurement in telemedicine rehabilitation based on intelligent wearable medical devices. IEEE Access 2019 Apr 25;7:54819-54823. [doi: 10.1109/access.2019.2913189]
7. Pimentel MAF, Redfern OC, Malycha J, Meredith P, Prytherch D, Briggs J, et al. Detecting deteriorating patients in the hospital: development and validation of a novel scoring system. Am J Respir Crit Care Med 2021 Jul 01;204(1):44-52. [doi: 10.1164/rccm.202007-2700oc]

8. Chi C, Ao S, Winkler A, Fu K, Xu J, Ho Y, et al. Predicting the mortality and readmission of in-hospital cardiac arrest patients with electronic health records: a machine learning approach. J Med Internet Res 2021 Sep 13;23(9):e27798 [FREE Full text] [doi: 10.2196/27798] [Medline: 34515639]
9. Moffat LM, Xu D. Accuracy of machine learning models to predict in-hospital cardiac arrest. Clin Nurse Spec 2022;36(1):29-44. [doi: 10.1097/nur.0000000000000644]
10. Pai MMM, Ganiga R, Pai RM, Sinha RK. Standard electronic health record (EHR) framework for Indian healthcare system. Health Serv Outcomes Res Method 2021 Jan 27;21(3):339-362. [doi: 10.1007/s10742-020-00238-0]
11. Murdoch B. Privacy and artificial intelligence: challenges for protecting health information in a new era. BMC Med Ethics 2021 Sep 15;22(1):122 [FREE Full text] [doi: 10.1186/s12910-021-00687-3] [Medline: 34525993]
12. Health Level 7 International. URL: https://www.hl7.org/ [accessed 2022-09-21]
13. FHIR v4.3.0. Health Level 7 International. URL: http://hl7.org/fhir/resourcelist.html [accessed 2022-09-21]
14. Observation vital signs panel profile. Health Level 7. URL: https://build.fhir.org/observation-vitalsigns.html [accessed 2022-09-21]
15. Pfaff ER, Champion J, Bradford RL, Clark M, Xu H, Fecho K, et al. Fast Healthcare Interoperability Resources (FHIR) as a meta model to integrate common data models: development of a tool and quantitative validation study. JMIR Med Inform 2019 Oct 16;7(4):e15199 [FREE Full text] [doi: 10.2196/15199] [Medline: 31621639]
16. Seong D, Jung S, Bae S, Chung J, Son D, Yi B. Fast Healthcare Interoperability Resources (FHIR)-based quality information exchange for clinical next-generation sequencing genomic testing: implementation study. J Med Internet Res 2021 Apr 28;23(4):e26261 [FREE Full text] [doi: 10.2196/26261] [Medline: 33908889]
17. Park HS, Kim KI, Chung H, Jeong S, Soh JY, Hyun YH, et al. A worker-centered personal health record app for workplace health promotion using national health care data sets: design and development study. JMIR Med Inform 2021 Aug 04;9(8):e29184 [FREE Full text] [doi: 10.2196/29184] [Medline: 34346894]
18. Chukwu E, Garg L, Obande-Ogbuinya N, Chattu VK. Standardizing primary health care referral data sets in Nigeria: practitioners' survey, form reviews, and profiling of Fast Healthcare Interoperability Resources (FHIR). JMIR Form Res 2022 Jul 07;6(7):e28510 [FREE Full text] [doi: 10.2196/28510] [Medline: 35797096]
19. Schleyer TKL, Rahurkar S, Baublet AM, Kochmann M, Ning X, Martin DK, FHIR Development Team, et al. Preliminary evaluation of the Chest Pain Dashboard, a FHIR-based approach for integrating health information exchange information directly into the clinical workflow. AMIA Jt Summits Transl Sci Proc 2019 May 06;2019:656-664 [FREE Full text] [Medline: 31259021]
20. HAPI-FHIR starter project. GitHub. URL: https://github.com/hapifhir/hapi-fhir-jpaserver-starter [accessed 2022-09-21]
21. Su C, Chiu S, Jang JR, Lai F. Improved inpatient deterioration detection in general wards by using time-series vital signs. Sci Rep 2022 Jul 13;12(1):11901 [FREE Full text] [doi: 10.1038/s41598-022-16195-2] [Medline: 35831415]
22. Ko H. Pseudonymization of healthcare data in South Korea. Nat Med 2022 Jan 17;28(1):15-16. [doi: 10.1038/s41591-021-01580-7] [Medline: 35039658]
23. Hung JC, Wang C. Exploring the website object layout of responsive web design: results of eye tracking evaluations. J Supercomput 2020 Apr 13;77(1):343-365. [doi: 10.1007/s11227-020-03283-1]
24. Kim J, Chae M, Chang H, Kim Y, Park E. Predicting cardiac arrest and respiratory failure using feasible artificial intelligence with simple trajectories of patient data. J Clin Med 2019 Aug 29;8(9):1336 [FREE Full text] [doi: 10.3390/jcm8091336] [Medline: 31470543]
25. Green M, Lander H, Snyder A, Hudson P, Churpek M, Edelson D. Comparison of the between the flags calling criteria to the MEWS, NEWS and the electronic Cardiac Arrest Risk Triage (eCART) score for the identification of deteriorating ward patients. Resuscitation 2018 Feb;123:86-91 [FREE Full text] [doi: 10.1016/j.resuscitation.2017.10.028] [Medline: 29169912]
26. Gulden C, Blasini R, Nassirian A, Stein A, Altun FB, Kirchner M, et al. Prototypical clinical trial registry based on Fast Healthcare Interoperability Resources (FHIR): design and implementation study. JMIR Med Inform 2021 Jan 12;9(1):e20470 [FREE Full text] [doi: 10.2196/20470] [Medline: 33433393]
27. Madrigal E, Le LP. Digital media archive for gross pathology images based on open-source tools and Fast Healthcare Interoperability Resources (FHIR). Mod Pathol 2021 Sep;34(9):1686-1695 [FREE Full text] [doi: 10.1038/s41379-021-00824-8] [Medline: 34035438]
28. González-Castro L, Cal-González VM, Del Fiol G, López-Nores M. CASIDE: a data model for interoperable cancer survivorship information based on FHIR. J Biomed Inform 2021 Dec;124:103953 [FREE Full text] [doi: 10.1016/j.jbi.2021.103953] [Medline: 34781009]

Abbreviations

API: application programming interface
EHR: electronic health record
ETL: Extract Transform Load
FHIR: Fast Healthcare Interoperability Resources

HIS: healthcare information system
IHCA: in-hospital cardiac arrest
PHR: personal health record

Edited by M Focsa; submitted 04.09.22; peer-reviewed by A Nassirian, R Saripalle, T Zhang; comments to author 20.09.22; revised version received 22.09.22; accepted 03.10.22; published 13.10.22.

Please cite as:
Tseng TW, Su CF, Lai F
Fast Healthcare Interoperability Resources for Inpatient Deterioration Detection With Time-Series Vital Signs: Design and Implementation Study
JMIR Med Inform 2022;10(10):e42429
URL: https://medinform.jmir.org/2022/10/e42429
doi:10.2196/42429
PMID:36227636

Original Paper

Tooth-Related Disease Detection System Based on Panoramic Images and Optimization Through Automation: Development Study

Changgyun Kim[1*], PhD; Hogul Jeong[1], DDS, PhD; Wonse Park[2*], DDS, PhD; Donghyun Kim[1*], MS

[1]AI Cloud R&D Center, InVisionLab Inc, Seoul, Republic of Korea

[2]Department of Advanced General Dentistry, College of Dentistry, Yonsei University & Institute for Innovation in Digital Healthcare, Seoul, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Hogul Jeong, DDS, PhD
AI Cloud R&D Center, InVisionLab Inc
G114, 128, Beobwon-ro, Songpa-gu
Seoul, 05854
Republic of Korea
Phone: 82 70 4415 2229
Fax: 82 50 4439 7765
Email: rari98@naver.com

Abstract

Background: Early detection of tooth-related diseases in patients plays a key role in maintaining their dental health and preventing future complications. Since dentists are not overly attentive to tooth-related diseases that may be difficult to judge visually, many patients miss timely treatment. The 5 representative tooth-related diseases, that is, coronal caries or defect, proximal caries, cervical caries or abrasion, periapical radiolucency, and residual root can be detected on panoramic images. In this study, a web service was constructed for the detection of these diseases on panoramic images in real time, which helped shorten the treatment planning time and reduce the probability of misdiagnosis.

Objective: This study designed a model to assess tooth-related diseases in panoramic images by using artificial intelligence in real time. This model can perform an auxiliary role in the diagnosis of tooth-related diseases by dentists and reduce the treatment planning time spent through telemedicine.

Methods: For learning the 5 tooth-related diseases, 10,000 panoramic images were modeled: 4206 coronal caries or defects, 4478 proximal caries, 6920 cervical caries or abrasion, 8290 periapical radiolucencies, and 1446 residual roots. To learn the model, the fast region-based convolutional network (Fast R-CNN), residual neural network (ResNet), and inception models were used. Learning about the 5 tooth-related diseases completely did not provide accurate information on the diseases because of indistinct features present in the panoramic pictures. Therefore, 1 detection model was applied to each tooth-related disease, and the models for each of the diseases were integrated to increase accuracy.

Results: The Fast R-CNN model showed the highest accuracy, with an accuracy of over 90%, in diagnosing the 5 tooth-related diseases. Thus, Fast R-CNN was selected as the final judgment model as it facilitated the real-time diagnosis of dental diseases that are difficult to judge visually from radiographs and images, thereby assisting the dentists in their treatment plans.

Conclusions: The Fast R-CNN model showed the highest accuracy in the real-time diagnosis of dental diseases and can therefore play an auxiliary role in shortening the treatment planning time after the dentists diagnose the tooth-related disease. In addition, by updating the captured panoramic images of patients on the web service developed in this study, we are looking forward to increasing the accuracy of diagnosing these 5 tooth-related diseases. The dental diagnosis system in this study takes 2 minutes for diagnosing 5 diseases in 1 panoramic image. Therefore, this system plays an effective role in setting a dental treatment schedule.

(JMIR Med Inform 2022;10(10):e38640) doi:10.2196/38640

KEYWORDS

object detection; tooth; diagnosis; panorama; dentistry; dental health; oral health; dental caries; image analysis; artificial intelligence; detection model; machine learning; automation; diagnosis system

Introduction

Usage of Medical Data and Artificial Intelligence in Health Care

Several recent studies [1-3] have used various medical data for eHealth care, but they are merely adding digital and network functions to the existing medical equipment, and remote services included in the treatment are unused. In addition, although eHealth care processes medical data and information through the networking function of doctors and patients, in reality, patients cannot obtain and confirm much information. Although a large amount of medical data has been accumulated, there has been a limit to using these data to provide information to patients and find new practical implications. As the importance of medical data has increased, a clinical data warehouse has been established to research how to utilize various medical data and for patients to find medical information easily through the provision of public and private medical data [4]. Various studies on the application of big data and artificial intelligence (AI) in medicine have shown that the University of North Carolina Healthcare has dramatically reduced the time and effort of medical staff by performing unstructured medical data analysis using content analytics and natural language processing and automatically extracting abnormal parts by machine reading and automatic processing algorithms in mammography screenings and pap smears [5]. Patients' conditions are diagnosed remotely after the initial treatment by clinical professionals providing them with the medical information to manage their disease [6]. Recently, a method that allows users to easily use various medical data based on their experiences and help them make decisions through optimal information delivery when applying it to medical systems has been studied [7]. Using medical data and AI, patients can prevent diseases in advance and increase their autonomy in treatment scheduling by receiving knowledge of their condition and medical information. In addition, AI using medical data can reduce medical time and cost by assisting doctors in treatment.

Dental Caries Diagnosis Using Images

Dental caries is diagnosed using videos and radiographs, and studies [8,9] have shown the processing of videos and images for a more accurate diagnosis of dental caries. In 2003, Møystad et al [10] diagnosed dental caries by using pre-enhanced Digora storage phosphor images while performing radiography on areas where tooth decay occurred and where panorama X-ray and computed tomography systems (Soredex Medical Systems) could not be used because of territorial issues. In 2017, Veena Divya et al [11] diagnosed dental caries by using the contrast map of a panoramic image, controlling the contrast of the bright and dark parts to make the blurred panoramic image clear. In the same year, Singh and Sehgal [12] added light contrast to panoramic images to enhance the clarity and diagnose dental caries by exploring the dark areas, which corresponded to dental caries in the images. In 2019, Kale et al [13] showed that mothers were able to diagnose dental caries in photos of normal and decayed teeth obtained with a smartphone by using an atlas. In 2020, the Laplacian filtering backpropagation algorithm was used to learn and diagnose dental caries [14]. In 2021, Bayraktar and Ayan [15] diagnosed dental caries by using image deep learning algorithms; that study used 1000 radiographic teeth data points for learning and validation. Labeling the dental caries was performed by a professional dentist, and dental caries in the premolars and molars were examined [15].

Importance of Dental Caries Diagnosis

Dental caries is one of the most common chronic diseases worldwide. Oral diseases are recognized as serious diseases like other systemic diseases and were classified by the World Health Organization in 2011 as serious noncommunicable diseases. The teeth are one of the most important organs in the body, and dental caries is one of the biggest causes of tooth disease [16]. Dental caries develop and progress in 4 stages, starting as a tiny black spot in stage 1, followed by enamel decay in stage 2, nerve damage in stage 3, and pulp damage and pus and inflammation in stage 4. Dental caries can be easily repaired with simple treatment in stages 1 and 2; however, if the initial stages 1 and 2 are not judged or are overlooked, dental caries progress to stages 3 and 4. This leads to complications such as toothache, inflammation, and acute osteomyelitis, which destroys the bones around the teeth. Therefore, it is important to prevent and manage dental caries. The management and early removal of dental caries through an initial diagnosis are essential factors for good dental health [17]. However, if there are no clinical symptoms in the early stages of dental caries, people often do not pay much attention. In addition, since dental treatment is generally performed to promptly resolve uncomfortable areas, dentists can also pass over without diagnosing any of the following: proximal caries, which occurs between teeth; periapical radiolucency, which occurs from the root apex; and residual root in the bone. Therefore, to solve this problem, AI can help dentists diagnose early dental caries and other tooth-related diseases that may be difficult to judge visually by using panoramic images. Through this system, dentists and patients can reduce treatment planning time and easily treat tooth problems before they worsen, and patients can identify problems with their teeth and improve their quality of life by preventing diseases that could occur in the future.

Although various simple and easy AI diagnostic methods in the dental field have been studied, there are limitations [18] in diagnosing dental caries accurately in tooth sections. Since previous models have been used for diagnosing dental caries in the entire tooth, there are limitations in diagnosing dental caries that require precise diagnosis, such as proximal and root caries. This study aims to learn and diagnose 5 tooth-related diseases (ie, coronal caries or defects, proximal caries, cervical caries or abrasion, periapical radiolucency, and residual root) by using image deep learning models, which can assist dentists' diagnosis by reducing treatment planning time.

Methods

Data Collection

Since this study evaluated 5 tooth-related diseases (ie, coronal caries or defects, proximal caries, cervical caries or abrasion, periapical radiolucency, and residual root), which are commonly diagnosed using dental imaging, training data were generated by collecting and labeling panoramic images with tooth-related diseases. This study used panoramic images provided by 50 dental clinics from 2001 to 2021. Data from 30 dental hospitals in Korea were collected, anonymized, and used for this study. Among the anonymized genders, there were 3702 males and 3783 females, with a total of 2515 unidentified persons who could not be identified. Population distribution by age group did not include teenagers; there were 1721 persons in their 20s, 956 persons in their 30s, 1134 persons in their 40s, 1351 persons in their 50s, 1914 persons in their 60s and older, and 2934 persons with unknown identities.

A total of 10,000 panoramic images with one or more of the following 5 tooth-related diseases were used for labeling: 4206 images of coronal caries or defects, 4478 images of proximal caries, 6920 images of cervical caries or abrasion, 8290 images of periapical radiolucency, and 1446 images of residual roots. As shown in Figure 1 and Table 1, coronal caries or defects showed defects or radiolucencies that lacked density compared to the normal in the coronal portion of the tooth, proximal caries showed radiolucency that lacked density compared to the normal in the adjacent surfaces between teeth, and cervical caries or abrasion showed radiolucency that lacked density compared to the normal in the cervical area of the tooth. In addition, periapical radiolucency showed a lower density than normal radiolucency in the periapical area, and residual root means that the coronal portion is completely lost and only the root portion remains. Each label was created by focusing on these findings on the panoramic images. We used 10,000 images of male and female Koreans to label each tooth-related disease. Radiologic specialists with over 20 years of dental imaging experience performed the labeling. It took 2 minutes on average for the radiologic specialists to read the 5 diseases presented in Table 1 on 1 panoramic image of the tooth, and it took approximately 6 hours on average to read 100, including the break time. Therefore, it took approximately 50 days to label 10,000 samples. Table 1 shows the standards agreed upon by the graders. This standard is presented in Oral Radiology: Principles and Interpretation [19].

Figure 1. Findings of each tooth-related disease (ie, coronal caries or defect, proximal caries, cervical caries or abrasion, periapical radiolucency, residual root, in clockwise order from the top left).

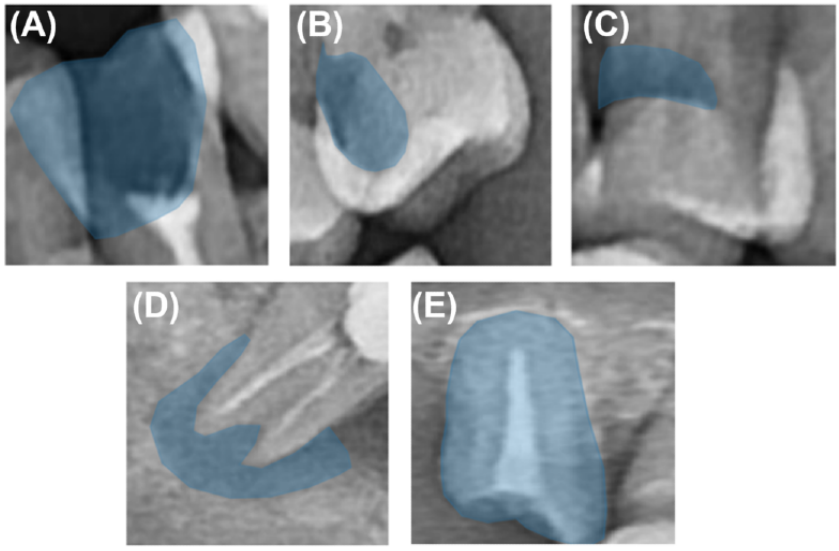


Table 1. Findings of each tooth-related disease.

Tooth-related diseases	Findings
Coronal caries or defect	Defect or radiolucency that lacks density compared to normal in the coronal portion of a tooth
Proximal caries	Radiolucency that lacks density compared to normal in the adjacent surfaces between teeth
Cervical caries or abrasion	Radiolucency that lacks density compared to normal in the cervical area of the tooth
Periapical radiolucency	Low density compared to normal in the periapical area of tooth
Residual root	Coronal portion of tooth is completely lost and only the root portion remains

Learning Model (Designing and Training the Model)

Labeling was performed by data collection and preprocessing, and thus, an image classification model was used to learn about each of the 5 tooth-related diseases. This study learned dental diseases by using fast region-based convolutional network (Fast R-CNN), residual neural network (ResNet), and inception. The model with the highest accuracy in disease detection was

selected. For training the model, 10,000 panoramic images were modeled in total: 4208 coronal caries or defects, 4478 proximal caries, 6920 cervical caries or abrasion, 8290 periapical radiolucency, and 1446 residual roots.

Model Used in This Study (Additional Case of Model Application)

Fast R-CNN has increased accuracy compared to the existing object detection algorithms because it extracts the image features and minimizes the noise in image analysis. Fast R-CNN consists of a convolution feature map and a region of interest feature vector [20]. The convolution feature map delivers the image to the convolution and max-pooling layers, and the received information is placed as a feature in the region of interest. Thereafter, the feature vector map is converted into a map with various features, and the object value of the object image of class K is determined by moving to the fully connected layers [21]. In this process, multiple work losses are minimized, and the learning accuracy is improved by using a loss function. Learning multiple classes of tooth-related diseases in 1 Fast R-CNN model sometimes results in errors in the detection of panoramic images with dark areas, as shown in Figure 2. Therefore, this study applies a single class to 1 Fast R-CNN model instead of multiple classes to improve the accuracy of detecting tooth-related diseases.

For image reading, a rectangular bounding box was first used, and segmentation was performed through an algorithm based on about 500 segmentation data. In the case of segmentation, accuracy was not calculated for the segmented data because it was used only for grasping the approximate accuracy. Thereafter, the coordinate values of the box-type tooth classes that are multilabeled in 1 tooth panoramic image were derived. Each disease corresponding to the derived coordinate value was classified by class. Then, each of the 5 tooth classes was applied to learning through the box coordinate values having the corresponding dental disease on the panoramic image. Through this, the input value for 1 model was constructed using the panoramic image data of 1 class and the box coordinate values corresponding to dental diseases. As shown in Figure 3, a bounding box was designated for each tooth-related disease, and the classes for each tooth-related disease were defined.

ResNet derives a value through the weight layer to solve the problem of overfitting owing to increased dimensional depth in deep learning, which adds the result learned through the previous weight layer to the activation function and delivers it to the next layer [22]. Therefore, this learning method, even if the depth of the learning layer deepens, solves the overfitting problem because important weights can be used for the next learning without forgetting the past learning results [23]. Because of these advantages, in this study, deep-layer learning is required to derive detailed results in learning panoramic images with similar image characteristics, and the ResNet model that can learn such a model was selected.

Inception, like ResNet, is created to solve the overfitting problem and the increase in computational traffic through a lot of learning when the size of the model is increased by increasing the depth of the layer [24]. In the inception model, it is possible to derive results in a fast learning time by using a small number of calculations, even in a model with a complex structure, by connecting only nodes with a high relationship between each node [25]. In addition, using various convolution filters, we derived a model that can make optimal judgments based on the features derived from each filter. This study evaluated 5 tooth-related diseases by using 3 models: Fast R-CNN, ResNet, and inception. To increase the detection accuracy for 5 tooth-related diseases, a model was designed through a process shown in Figure 4 (additional model), and the 5 tooth-related diseases were learned through Fast R-CNN, ResNet, and inception. In learning tooth-related disease data (the result of the additional model), the 3 models provided good performance for multi-class learning. However, for each part of the panoramic image composed of the contrast ratio of white and black, if multiple classes are learned in one detection model for tooth-related diseases that have similar characteristics but different sizes, there were cases where the black background was detected as a tooth-related disease. As the learning proceeded by inputting data for 5 tooth-related diseases as a whole, more black screens were learned, and the results are shown in Figure 2. As shown in the box in Figure 2, there are cases where areas such as the background of other panoramic images that are not included in the teeth are detected. To solve the problem of multi-class learning, as shown in Figure 2, professional reading experts labeled 10,000 images in a bounding box form with 5 dental diseases in a single tooth image and finally converted it into the CSV format. Label information corresponding to each dental disease was extracted from the data set containing the labeling information of 5 dental diseases, and each data set was derived for each of the 5 dental diseases. Therefore, 5 CSV-format data sets that were composed of panoramic images were modeled in total: 4208 coronal caries or defects, 4478 proximal caries, 6920 cervical caries or abrasion, 8290 periapical radiolucency, and 1446 residual roots. Further, depending on the model, DICOM (digital imaging and communications in medicine) to BMP (bitmap) conversion was performed, and auto brightness correction and adjustment were partially performed.

Figure 2. Case of misclassification of tooth-related diseases. The green boxes represent detected areas that are not included in the teeth.

Figure 3. Bounding box for each tooth-related disease.

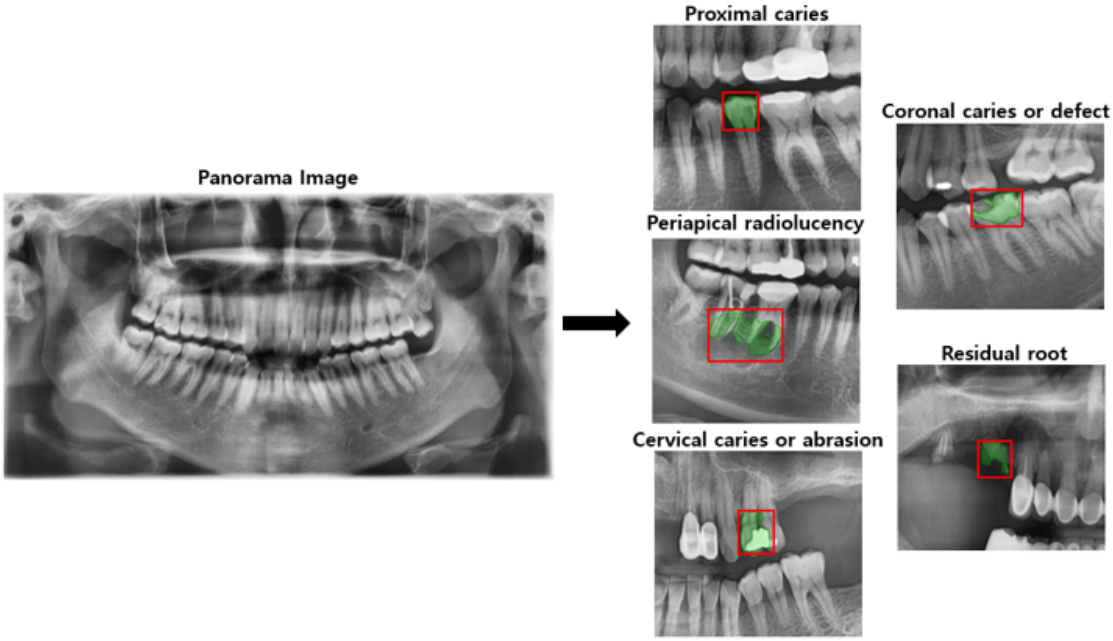
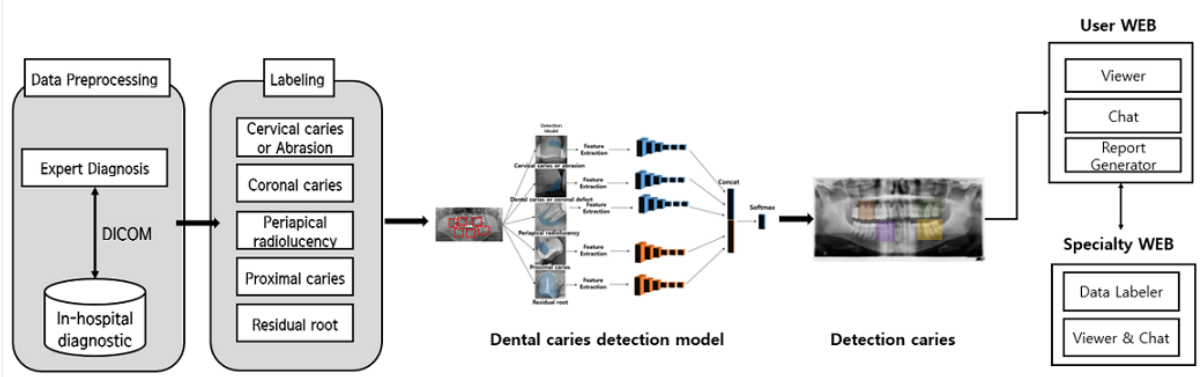


Figure 4. Integrated detection system for the learning process. DICOM: digital imaging and communications in medicine.



Development of an Integrated Detection System for Tooth-Related Diseases

While learning about the 5 tooth-related diseases through Fast R-CNN, ResNet, and inception, problems, as shown in Figure 2, appeared. To solve these problems, a single model was applied to 1 class of tooth-related diseases to create a training model for each of the 5 tooth-related diseases so that varying locations and sizes of the diseases could be detected in detail.

Based on the process shown in Figure 4, 5 tooth-related diseases were learned, and dentists and experts designed a real-time diagnosis, as shown in Figure 5. In designing the process, the service and administrator web were implemented using Python (version 3.6)-based Flask [26] engines (version 1.0.0), and the web page configuration was implemented using Jinja template-based HTML and Vanilla JavaScript [27]. The communication part of the AI application programming interface was composed of a Python-based Flask engine, which was installed within the Flask engine through model learning using TensorFlow 2.0.0 (Google Brain Team). Additionally, the image data of the database server were divided into file name, photographing date, patient name, patient age, image labeling prediction model data, and image labeling correct answer data to assist the dentist in the diagnosis. In the form of training/validation/test, splits were first performed and then labeled. A total of 6000 pieces were used for training, 2000 pieces were used for validation, and the remaining 2000 pieces were used for test splitting. In fact, we used ResNet/inception as the backbone of Fast R-CNN. As the input value for one model, learning data were constructed using the panoramic image data of one class and the box coordinate values corresponding to dental diseases. Through this, the input value for one model was constructed using the panoramic image data of one class and the box coordinate values corresponding to dental diseases.

In the training layer structure of each model of Fast R-CNN, ResNet, and inception, looking at the structure of the Fast R-CNN model (region proposal → CNN classification → region of interest projection), region of interest projection and bounding box regression were performed through region of interest pooling. The model is configured as shown in Figure 5, and 300 range boxes for each dental disease were specified using the CNN model in the region proposal for dental disease detection, and the features of the range corresponding to a specific class were identified. At this time, after converting features of fixed sizes in the region of interest pooling layer into a feature map, a feature vector was generated with a fully connected layer corresponding to each feature. At this time, for each feature, the position of the corresponding class was predicted using SoftMax and Bbox regressor. The epoch of model training was performed 100,000 times, and the learning rate was set to 0.001.

ResNet improves the accuracy by reducing the depth of the learning layer and increasing the performance compared to the CNN model, which is an existing image analysis model, through residual learning. In order to increase the learning accuracy in general CNNs, many layers are stacked. However, such a deep layer can lower the accuracy of the learning model. When learning through residual learning, the positive error rate can be lowered even when learning in a deep layer. When ResNet derives a value from the weight layer through the activation function in the convolution operation, it imports the previously learned information as it is, as shown in Figure 5, and learns the residual information, F(x). Looking at the formula, when the input value x is input, the first weight value is multiplied, and the activation function is multiplied by the second weight value. At this time, it is additionally multiplied by x identity, that is, x value. Therefore, since the result is derived through continuous repetition of this process, y is derived by adding a multiple convolutional layer $F(x,\{W_i\})$ and short connection $W_s x$, which takes the existing input value as it is, to the result value.

$$y = F(x,\{Wi\}) + Wsx$$

In this way, by adding information to the result derived from the weight layer, information can be added and computational complexity can be reduced so that a model with faster learning and better performance can be derived. Since ResNet learns 1 dental disease by using 5 models as 1 model, 50 hidden layers of each dental disease were designated for learning. For training, like Fast R-CNN, the training epoch was performed 100,000 times, and the learning rate was set to 0.001.

The inception model connects the highly correlated nodes when the correlation between each node is high in the fully connected architecture and does not connect the rest so that N clusters are created for each feature. When creating a connected architecture, we additionally convolve features that are far from each other through filters of various sizes for nonuniform and inefficient sparse structures and reduce the number of channels by using a 1 × 1 filter for nodes with high correlation. The inception model was constructed, as shown in Figure 6. For the model configuration, a dental disease detection model was built using 10 pooling layers. The training epoch was performed 100,000 times, and the learning rate was set to 0.001. When a list of images is received from a computer connected to the X-ray equipment and the data are stored in the server database, a separate image is retransmitted to a system that is requested to be read from the stored data. Thereafter, it provides information read through a detection model for tooth-related diseases in real time so that it can assist dentists in shortening the reading time.

The overall flow diagram is shown in Figure 6 and is divided into service, manager, and AI algorithm categories. In the service web, the data for each tooth-related disease previously labeled by experts and the updated panoramic images are continuously accumulated and provided to the server. In the manager app, the accumulated data are transmitted to the server, and the transmitted panoramic image is read by dental experts to determine the tooth-related disease. Then, the analysis data are collected through labeling, and the collected data are used to derive the result by using an AI algorithm.

Based on the process shown in Figure 5, the detailed process of the tooth-related disease determination system proposed in this study was constructed, and it can be divided into 3 parts (service, system, and personal computer). The service part is designed to receive panoramic image data and read information through the website, and the messaging system is designed for users to communicate through the channel talk application

programming interface [23]. The information provided to the readers was labeled so that the AI model could be learned, and it was designed to enable continuous data updates. In addition, when the labeling result was applied to the AI model and the AI result was judged again by the reader, it was updated to Case 1 if it was correctly judged and to Case 2 when the judgment was incorrect. Therefore, after being read accurately again by the reader, the accuracy of the model was improved through continuous data updates with the AI server. In the system, a server was built to enable the website of the service part to work. The server was built based on Flask, and it was largely divided into the presentation, business, and persistence layers [28,29]. The server connects the user and client systems through 3 layers and enables the movement of data in the database. The database was designed using MongoDB [30], which can quickly operate various types of data. AI, chatting, image, and message servers were built into MongoDB to increase the real-time movement speed of the data. The AI server, which plays a role in providing tooth-related disease reading results, updates the results of expert reading provided by doctors and provides the doctor with tooth-related disease results on new images to improve accuracy through mutual feedback, which helps users to understand by providing feedback on the opinions of users on the personal computer. Finally, it stores the dental panoramic image provided through the image server or provides medical information to personal computer users so that they can view and continuously manage the medical records whenever necessary.

Figure 5. Flow diagram of the learning process. AI: artificial intelligence; PC: personal computer.

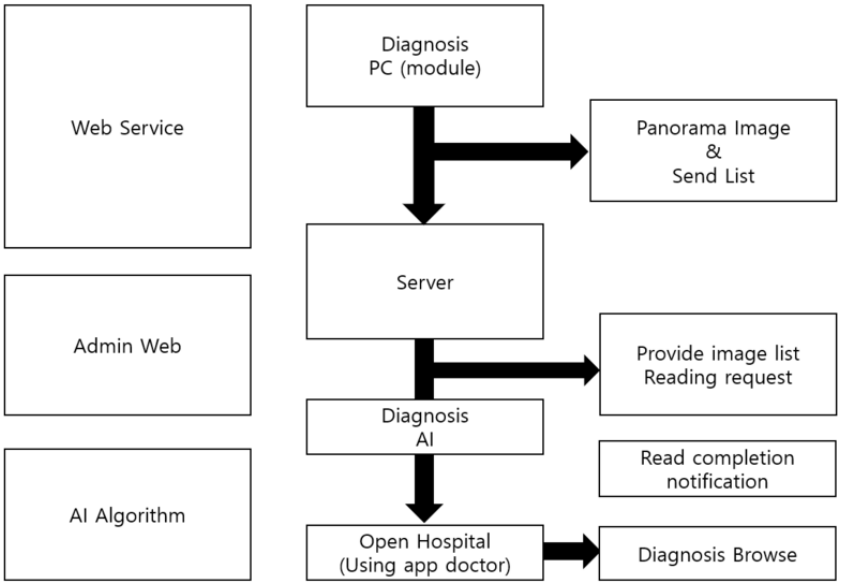
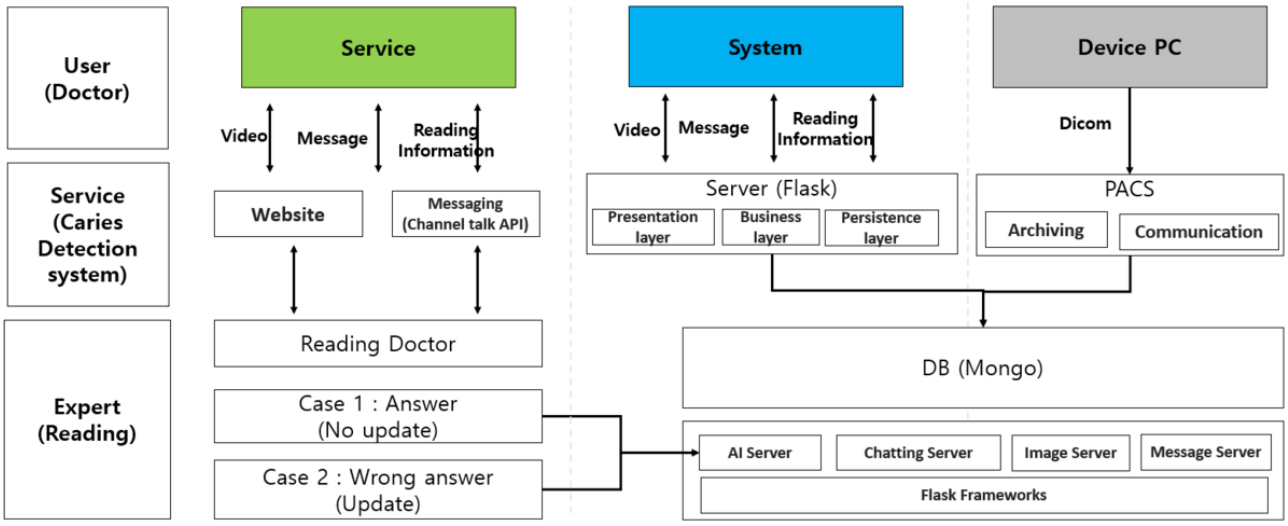


Figure 6. Schematic diagram of a detection system for tooth-related diseases. AI: artificial intelligence; API: application programming interface; DB: database; PACS: picture archiving and communication system; PC: personal computer.



Ethical Considerations

Since the data is a retrospective study, it was processed in the direction of protecting personal information through database anonymization, etc. In addition, data collected for research purposes were collected through Cheongju University Bioethics Committee IRB (1041107-202208-HR-024-01).

Results

Detection System

The web service in this study was built based on the process presented in Figure 6 and Figure 7. As shown in Figure 7, panoramic images facilitate faster judgment of dental-related diseases than the conventional doctor's diagnosis techniques. The detection system aids and shortens the treatment time through the transmission of images taken in real time. Figure 7 shows a case in which a tooth disease was correctly judged and another case in which a dental disease was incorrectly judged. Since cases can be judged inaccurately, doctors can use this auxiliary system to check the patient's condition once again. Figure 8 shows that patients can check their panoramic images on the web, and they can know about the treatment plan and receive information on tooth-related diseases for effective disease management.

Figure 7. Success and fail cases in the detection of the 5 dental diseases.

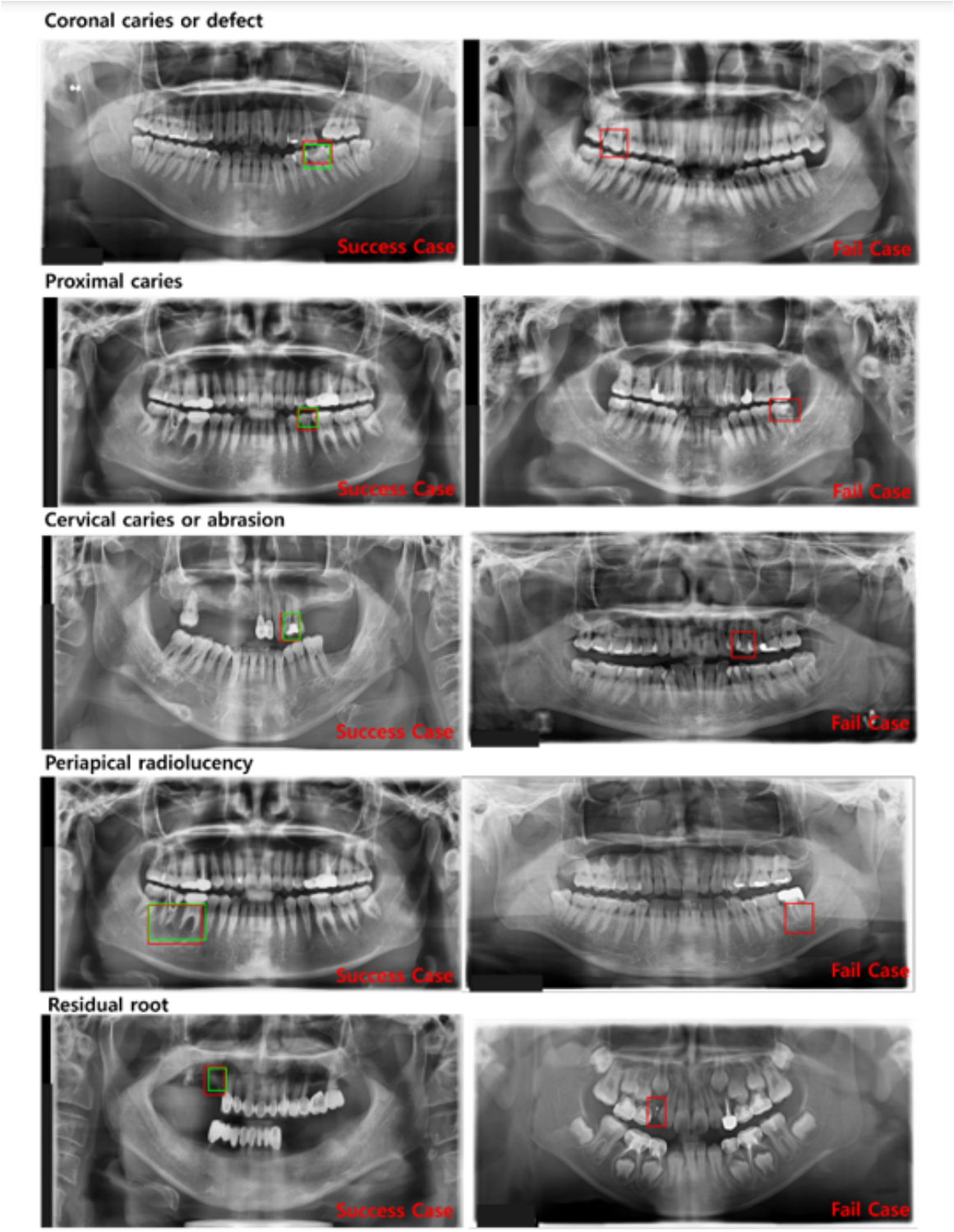
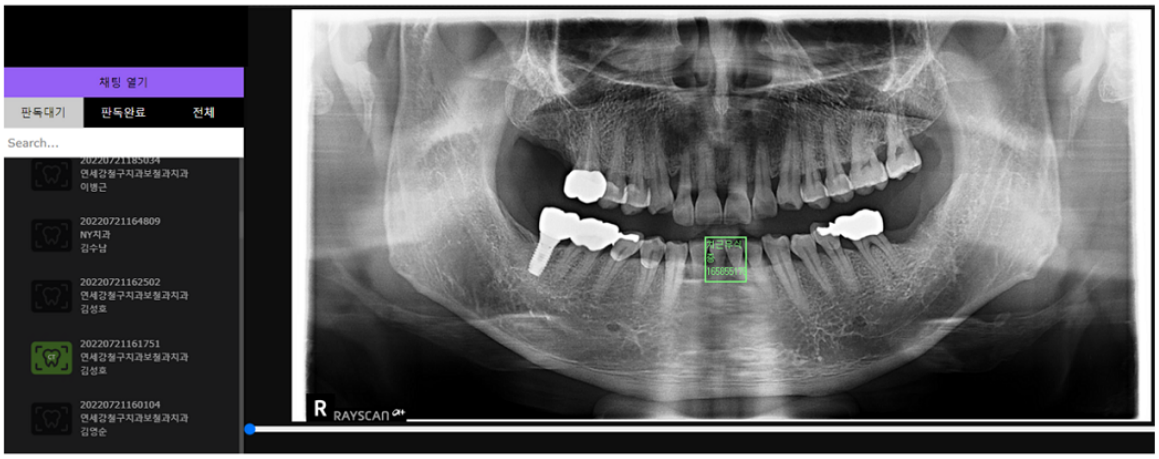


Figure 8. Web system for tooth-related diseases.



Benefits of Using Web-Based Systems and User Interface

To detect tooth-related diseases, Fast R-CNN, which has the best performance among the classification models from the point of view of a dentist, was applied. For learning the model for judging the 5 types of tooth-related diseases, dentists can update the panoramic images in real time through the web application programming interface and continuously collect data by improving the accuracy through additional updates on the tooth-related disease detection labeling. Figure 6 shows how the tooth-related disease judgment web service screen appears. Figure 6 shows that by providing doctors and patients with the diagnosis of their diseases through the patient's panoramic image, past medical records, and current status on the web, doctors can provide prompt treatment for dental diseases and patients can monitor their dental status. Therefore, from the patient's perspective, users can check medical records and treatment areas through the web screen of the panoramic image provided by the hospital where they have been treated and check for tooth-related diseases. In addition, because the treatment time and the subsequent treatment times can be known, users can use this system to manage their tooth-related diseases, which require continuous management.

Model Comparison Results

This study created a detection model for 5 dental diseases that are difficult to judge visually (ie, coronal caries or defect, proximal caries, cervical caries or abrasion, periapical radiolucency, and residual root) by using a dental panoramic image. Fast R-CNN, ResNet, and inception have previously been used to learn about dental disease detection [20,22,31]. In training the model, 4206 cases of coronal caries or defects, 4478 cases of proximal caries, 6920 cases of cervical caries or abrasion, and 8290 cases of periapical radiolucency, and 1446 cases of residual roots were trained among a total of 10,000 panoramic images. Therefore, a model for judging the 5 types of dental caries using 1 panoramic image was developed by creating a training model for each dental disease into one detection model through an integrated detection system for dental diseases. Regarding the number of training sessions, all 3 models were trained 200,000 times, the results were compared, and the model with the highest accuracy was selected. The results of deriving the precision, sensitivity, and specificity of the detection results for the 5 dental diseases are shown in Figure 8. As shown in Figure 8, the coronal defect showed the highest specificity, with an average specificity of 90 or more. In addition, the sensitivity was found to be above 80 on average, indicating that it would show high accuracy even when other data were used for learning.

Table 2 shows the results of learning with Fast R-CNN, ResNet, and inception for the 5 tooth-related diseases. As shown in Table 2, 5 tooth-related diseases were detected with an average accuracy of over 90%. Also, as shown in Figure 6, the specificity is the highest for the 5 tooth-related diseases. This means that each tooth-related disease can be detected with high accuracy. With the tooth-related disease detection web service presented in this study, considerable time can be saved in diagnosing tooth-related diseases. On average, it takes about 1 minute for dental doctors to judge 5 dental diseases on 1 panoramic image. However, if the system proposed in this study is used, the results of the classification model can be judged at once through the user interface, and the time can be reduced to about 10 seconds in judging dental diseases. Therefore, it is judged to be an effective system to assist in the judgment of dental diseases.

Table 2. Tooth-related disease detection results.

Model, diseases	Precision	Sensitivity	Specificity
Fast region-based convolutional network			
Coronal caries or defect	0.785	0.708	0.982
Proximal caries	0.484	0.792	0.918
Cervical caries or abrasion	0.795	0.767	0.952
Periapical radiolucency	0.824	0.953	0.895
Residual root	0.640	0.904	0.972
Inception			
Coronal caries or defect	0.253	0.609	0.848
Proximal caries	0.327	0.783	0.883
Cervical caries or abrasion	0.444	0.707	0.785
Periapical radiolucency	0.371	0.946	0.556
Residual root	0.232	0.893	0.873
Residual neural network			
Coronal caries or defect	0.2101	0.395	0.876
Proximal caries	0.685	0.377	0.987
Cervical caries or abrasion	0.378	0.011	0.996
Periapical radiolucency	0.308	0.883	0.451
Residual root	0.225	0.744	0.89

Discussion

Strengths and Limitations

This study has several advantages. The use of panoramic images of individual patients in dentistry is a complex procedure. This study designed a model that could determine 5 types of dental caries by acquiring various panoramic image data and collecting 10,000 pieces of data with various oral structures and dental caries. Therefore, a tooth-related disease determination system with high accuracy and without complex procedures was developed. However, since there is a large deviation in the number of classes for each tooth-related disease, there was a problem in that the learning accuracy was slightly lowered where the number of analysis groups was small. The accuracy of the model is expected to be improved by collecting and supplementing data through continuous updates by using real-time panoramic images uploaded to the web.

Conclusions

In this study, the tooth-related disease judgment system identified 5 types of tooth-related diseases that are difficult to determine clinically (visually) by using an AI model, and this information was provided on the web to create a system that allows doctors and patients to make real-time judgments. The trained model labeled 5 dental caries through 10,000 panoramic images. Accuracy was compared using Fast R-CNN, ResNet, and inception models, which are good models for detection. Among these models, Fast R-CNN was finally used, which has the highest accuracy. Therefore, Fast R-CNN can be used to shorten the time required for the diagnosis and treatment of dental caries. In addition, by updating the captured panoramic images of patients on the web service developed in this study, the system can acquire new data and further increase the accuracy of diagnosing tooth-related diseases. Additionally, the patient can be aware of the tooth areas where he or she has received treatment, the treatment time, and the type of caries, so that he or she can adjust the schedule for the future dental visit, which will aid in continuous management of dental health. Thus, this study is meaningful as it collects learning data from cases embodied as actual services and implements a prototype-type service based on the collected data. In the future, it will be possible to develop a model for predicting overall oral diseases with panoramic images through additional learning of various dental diseases.

Conflicts of Interest

None declared.

References

1. Dash S, Shakyawar S, Sharma M, Kaushik S. Big data in healthcare: management, analysis and future prospects. J Big Data 2019 Jun 19;6(1):54. [doi: 10.1186/s40537-019-0217-0]

2. Huang H, Gong T, Ye N, Wang R, Dou Y. Private and Secured Medical Data Transmission and Analysis for Wireless Sensing Healthcare System. IEEE Trans. Ind. Inf 2017 Jun;13(3):1227-1237. [doi: 10.1109/TII.2017.2687618]
3. Pirbhulal S, Samuel OW, Wu W, Sangaiah AK, Li G. A joint resource-aware and medical data security framework for wearable healthcare systems. Future Generation Computer Systems 2019 Jun;95:382-391. [doi: 10.1016/j.future.2019.01.008]
4. Garcelon N, Neuraz A, Benoit V, Salomon R, Kracker S, Suarez F, et al. Finding patients using similarity measures in a rare diseases-oriented clinical data warehouse: Dr. Warehouse and the needle in the needle stack. J Biomed Inform 2017 Sep;73:51-61 [FREE Full text] [doi: 10.1016/j.jbi.2017.07.016] [Medline: 28754522]
5. Seong J, Song W. Assessment of innovation policy coordination through the Korean Office of Science, Technology and Innovation. Korea Science. 2013. URL: http://koreascience.or.kr/article/JAKO201354447932240.jsp%3Fkj=KJKHCF&py=2004&vnc=v7n4&sp=251 [accessed 2022-10-12]
6. Dixon PM, Tremaine WH, Pickles K, Kuhns L, Hawe C, McCann J, et al. Equine dental disease part 1: a long-term study of 400 cases: disorders of incisor, canine and first premolar teeth. Equine Vet J 1999 Sep;31(5):369-377. [doi: 10.1111/j.2042-3306.1999.tb03835.x] [Medline: 10505951]
7. Barbazza E, Klazinga NS, Kringos DS. Exploring the actionability of healthcare performance indicators for quality of care: a qualitative analysis of the literature, expert opinion and user experience. BMJ Qual Saf 2021 Dec;30(12):1010-1020 [FREE Full text] [doi: 10.1136/bmjqs-2020-011247] [Medline: 33963072]
8. Prados-Privado M, García Villalón J, Martínez-Martínez CH, Ivorra C, Prados-Frutos JC. Dental Caries Diagnosis and Detection Using Neural Networks: A Systematic Review. J Clin Med 2020 Nov 06;9(11):3579 [FREE Full text] [doi: 10.3390/jcm9113579] [Medline: 33172056]
9. Lee JH, Kim DH, Jeong SN, Choi SH. Detection and diagnosis of dental caries using a deep learning-based convolutional neural network algorithm. J Dent 2018 Oct;77:106-111. [doi: 10.1016/j.jdent.2018.07.015] [Medline: 30056118]
10. Møystad A, Svanaes DB, van der Stelt P, Gröndahl HG, Wenzel A, van Ginkel F, et al. Comparison of standard and task-specific enhancement of Digora storage phosphor images for approximal caries diagnosis. Dentomaxillofac Radiol 2003 Nov;32(6):390-396. [doi: 10.1259/dmfr/76382099] [Medline: 15070842]
11. Veena Divya K, Jatti A, Joshi R, et al. Characterization of dental pathologies using digital panoramic X-ray images based on texture analysis. 2017 Jul 11 Presented at: Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); 2017; Jeju Island p. 592-595. [doi: 10.1109/embc.2017.8036894]
12. Singh P, Sehgal P. Automated caries detection based on radon transformation and DCT. 2017 Presented at: International Conference on Computing, Communication and Networking Technologies (ICCCN); July 31; Vancouver, Canada p. 1-6. [doi: 10.1109/icccnt.2017.8204030]
13. Kale S, Kakodkar P, Shetiya S. Assessment of mother's ability in caries diagnosis, utilizing the smartphone photographic method. J Indian Soc Pedod Prev Dent 2019;37(4):360. [doi: 10.4103/jisppd.jisppd_349_18]
14. Geetha V, Aprameya KS, Hinduja DM. Dental caries diagnosis in digital radiographs using back-propagation neural network. Health Inf Sci Syst 2020 Dec;8(1):8 [FREE Full text] [doi: 10.1007/s13755-019-0096-y] [Medline: 31949895]
15. Bayraktar Y, Ayan E. Diagnosis of interproximal caries lesions with deep convolutional neural network in digital bitewing radiographs. Clin Oral Investig 2022 Jan;26(1):623-632 [FREE Full text] [doi: 10.1007/s00784-021-04040-1] [Medline: 34173051]
16. Kim Y, Lee JH, et al. Evaluation of national health insurance coverage of periodontal scaling: A nationwide cohort study in Korea. Journal of the Korean Dental Association. 2016. URL: https://koreascience.kr/article/JAKO201664558247211.page [accessed 2022-10-12]
17. Kim D, Cho H. Comparative Analysis of Dental Caries Detection Technologies based on Computer-aided Diagnosis System. KIEE 2019 Feb 28;68(2):350-358. [doi: 10.5370/kiee.2019.68.2.350]
18. Ezhov M, Gusarev M, Golitsyna M, Yates JM, Kushnerev E, Tamimi D, et al. Clinically applicable artificial intelligence system for dental diagnosis with CBCT. Sci Rep 2021 Jul 22;11(1):15006 [FREE Full text] [doi: 10.1038/s41598-021-94093-9] [Medline: 34294759]
19. White SC, Pharaoh MJ. Oral Radiology-E-Book: Principles and Interpretation. St Louis, Missouri: Elsevier Health Sciences; 1982.
20. Girshick R. Fast R-CNN. 2015 Presented at: 2015 IEEE International Conference on Computer Vision (ICCV); December 7-13; Santiago, Chile. [doi: 10.1109/ICCV.2015.169]
21. Ray PP, Dash D, De D. Edge computing for Internet of Things: A survey, e-healthcare case study and future direction. Journal of Network and Computer Applications 2019 Aug;140:1-22. [doi: 10.1016/j.jnca.2019.05.005]
22. Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, inception-ResNet and the impact of residual connections on learning. 2017 Presented at: AAAI'17: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence; February 4-9; San Francisco, California. [doi: 10.1609/aaai.v31i1.11231]
23. Poggio T, Mhaskar H, Rosasco L, Miranda B, Liao Q. Why and when can deep-but not shallow-networks avoid the curse of dimensionality: A review. Int. J. Autom. Comput 2017 Mar 14;14(5):503-519. [doi: 10.1007/s11633-017-1054-2]
24. Szegedy C. Rethinking the inception architecture for computer vision. 2016 Jun 27 Presented at: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016; Las Vegas, NV, USA. [doi: 10.1109/cvpr.2016.308]

25. Guan Q, Wan X, Lu H, Ping B, Li D, Wang L, et al. Deep convolutional neural network Inception-v3 model for differential diagnosing of lymph node in cytological images: a pilot study. Ann Transl Med 2019 Jul;7(14):307 [FREE Full text] [doi: 10.21037/atm.2019.06.29] [Medline: 31475177]
26. Flanagan D. Java-Script: The Definitive Guide. URL: https://pepa.holla.cz/wp-content/uploads/2016/08/JavaScript-The-Definitive-Guide-6th-Edition.pdf [accessed 2022-10-12]
27. Wang H, Prendinger H, Igarashi T. Communicating emotions in online chat using physiological sensors and animated text. Human Factors in Computing Systems 2004:1171. [doi: 10.1145/985921.986016]
28. Zhou Z, Chen Z. Performance evaluation of transparent persistence layer in Java applications. 2010 Oct 10 Presented at: International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery; October 10; Huangshan, China p. 21-26. [doi: 10.1109/cyberc.2010.15]
29. Saimi A, Symora T, et al. Presentation layer framework of web application systems with server-side Java technology. 2010 Oct 06 Presented at: Proceedings of the 24th Annual International Computer Software and Applications Conference; October 6; Taipei, Taiwan. [doi: 10.1109/cmpsac.2000.884769]
30. Boicea A, Radulescu F, et al. MongoDB vs Oracle: database comparison. 2012 Presented at: Third International Conference on Emerging Intelligent Data and Web Technologies; September 21; Bucharest, Romania p. 330-335. [doi: 10.1109/eidwt.2012.32]
31. Resnet in Resnet: generalizing residual architectures. Deep AI. URL: https://deepai.org/publication/resnet-in-resnet-generalizing-residual-architectures [accessed 2022-10-12]

Abbreviations

AI: artificial intelligence
Fast R-CNN: fast region-based convolutional network
ResNet: residual neural network

Edited by C Lovis, J Hefner; submitted 25.04.22; peer-reviewed by G Lim, Z Li; comments to author 16.05.22; revised version received 11.07.22; accepted 11.08.22; published 31.10.22.

Please cite as:
Kim C, Jeong H, Park W, Kim D
Tooth-Related Disease Detection System Based on Panoramic Images and Optimization Through Automation: Development Study
JMIR Med Inform 2022;10(10):e38640
URL: https://medinform.jmir.org/2022/10/e38640
doi:10.2196/38640
PMID:36315222

Original Paper

Coronary Artery Computed Tomography Angiography for Preventing Cardio-Cerebrovascular Disease: Observational Cohort Study Using the Observational Health Data Sciences and Informatics' Common Data Model

Woo Kyung Bae[1*], MPH, MD; Jihoon Cho[2*], MS; Seok Kim[2], MPH; Borham Kim[2], BSN, RN; Hyunyoung Baek[2], MPH, RN; Wongeun Song[2], MS; Sooyoung Yoo[2], PhD

[1]Department of Family Medicine, Health Promotion Center, Seoul National University Bundang Hospital, Republic of Korea, Bundang-gu, Seongnam-si, Gyeonggi-do, Republic of Korea

[2]Healthcare Information and Communication Technology Research Center, Office of eHealth Research and Business, Seoul National University Bundang Hospital, Republic of Korea, Seongnam-si, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Sooyoung Yoo, PhD
Healthcare Information and Communication Technology Research Center, Office of eHealth Research and Business
Seoul National University Bundang Hospital
Republic of Korea
172, Dolma-ro, Bundang-gu
Seongnam-si, 13605
Republic of Korea
Phone: 82 010 9053 7094
Email: yoosoo0@snubh.org

Abstract

Background: Cardio-cerebrovascular diseases (CVDs) result in 17.5 million deaths annually worldwide, accounting for 46.2% of noncommunicable causes of death, and are the leading cause of death, followed by cancer, respiratory disease, and diabetes mellitus. Coronary artery computed tomography angiography (CCTA), which detects calcification in the coronary arteries, can be used to detect asymptomatic but serious vascular disease. It allows for noninvasive and quick testing despite involving radiation exposure.

Objective: The objective of our study was to investigate the effectiveness of CCTA screening on CVD outcomes by using the Observational Health Data Sciences and Informatics' Observational Medical Outcomes Partnership Common Data Model (OMOP-CDM) data and the population-level estimation method.

Methods: Using electronic health record–based OMOP-CDM data, including health questionnaire responses, adults (aged 30-74 years) without a history of CVD were selected, and 5-year CVD outcomes were compared between patients undergoing CCTA (target group) and a comparison group via 1:1 propensity score matching. Participants were stratified into low-risk and high-risk groups based on the American College of Cardiology/American Heart Association atherosclerotic cardiovascular disease (ASCVD) risk score and Framingham risk score (FRS) for subgroup analyses.

Results: The 2-year and 5-year risk scores were compared as secondary outcomes between the two groups. In total, 8787 participants were included in both the target group and comparison group. No significant differences (calibration P=.37) were found between the hazard ratios of the groups at 5 years. The subgroup analysis also revealed no significant differences between the ASCVD risk scores and FRSs of the groups at 5 years (ASCVD risk score: P=.97; FRS: P=.85). However, the CCTA group showed a significantly lower increase in risk scores at 2 years (ASCVD risk score: P=.03; FRS: P=.02).

Conclusions: Although we could not confirm a significant difference in the preventive effects of CCTA screening for CVDs over a long period of 5 years, it may have a beneficial effect on risk score management over 2 years.

(JMIR Med Inform 2022;10(10):e41503) doi:10.2196/41503

KEYWORDS

cardiovascular diseases; coronary artery computed tomography angiography; observational study; common data model; population level estimation; cardiology; vascular disease; medical informatics; computed tomography; angiography; electronic health record; risk score; health data science; data modeling

Introduction

Cardio-cerebrovascular diseases (CVDs) result in 17.5 million deaths annually worldwide, accounting for 46.2% of noncommunicable causes of death, and are the leading cause of death, followed by cancer, respiratory disease, and diabetes mellitus [1]. CVDs involve demographic factors (age, sex, and family history), pre-existing conditions (hypertension, diabetes mellitus, and hyperlipidemia), and lifestyle and environmental factors. Unlike demographic characteristics, lifestyle factors, such as an inappropriate diet, a lack of exercise, smoking, stress, and excessive drinking, can be improved to reduce the risk of CVDs [2].

Coronary artery computed tomography angiography (CCTA) detects calcification in the coronary arteries and can be used to detect asymptomatic but serious vascular disease. It allows for noninvasive and quick testing despite involving radiation exposure [3,4]. For these reasons, many studies have investigated the early detection of CVDs by using CCTA, which enables prompt treatment and results in better outcomes.

In recent years, there has been debate about whether screening via CCTA helps prevent CVDs in populations with varying degrees of risk. CCTA has been recommended to predict CVDs in patients with cancer [2,5], but among asymptomatic individuals, the evidence about its effectiveness is inconsistent.

We aimed to study the effectiveness of CCTA screening by analyzing observational health checkup data from electronic health records (EHRs) in the form of the Observational Medical Outcomes Partnership Common Data Model (OMOP-CDM), using a cohort study design [6]. The OMOP-CDM standardizes disparate data and enables the analysis of deidentified, large-scale observational data in a distributed research data network. Moreover, as the data are standardized, the same analytical codes can be used to conduct efficient analyses through the data network. Observational Health Data Sciences and Informatics (OHDSI)—an open international collaborative community—provides an open-source analytics tool for OMOP-CDM data that produces scientific, reliable, and reproducible evidence.

Using the OHDSI analytics tool, we performed a comparative effectiveness study of CVD outcomes in asymptomatic patients without a history of CVD who underwent a health checkup at a tertiary university hospital. The conventional assessments of CVD risk, namely assessments of the Framingham risk score (FRS) and the American College of Cardiology/American Heart Association (ACC/AHA) atherosclerotic cardiovascular disease (ASCVD) risk score, were used to stratify the participants into high-risk and low-risk groups for stratified analyses. Although the risk of CVD increases with age, we compared differences between the two groups after 2 and 5 years to assess the short-term benefits of CCTA-based screening and whether it can help prevent CVDs.

Methods

Data Sources

The study site was the Seoul National University Bundang Hospital (SNUBH), which is located in the Seoul metropolitan area. The SNUBH collected OMOP-CDM version 5.3 data based on EHRs from 2003 to 2020. The data included patients' demographic information, clinical information (diagnoses, medications, tests, surgeries and procedures, family histories, past histories, and nursing flowcharts), and health questionnaire responses. The health questionnaire responses about medical history, family history, socioeconomic status, medication history, marital status, exercise and physical activity status, and depression assessment results were converted to OMOP-CDM data. In this study, we used the deidentified OMOP-CDM data that the SNUBH collected from over 2 million patients, including outpatients, inpatients, and emergency department visits.

Ethical Considerations

This study adhered to the relevant guidelines and regulations of the SNUBH Institutional Review Board (IRB). As the OMOP-CDM is a deidentified data set, the study was exempted from review by the SNUBH IRB (IRB number: X-2202-736-903).

Study Design

This was a retrospective, observational, comparative cohort study that used OMOP-CDM–formatted EHR data. We analyzed data from adults aged 30 to 74 years who underwent a health checkup between April 1, 2003, and December 31, 2015, and were followed up for at least 5 years. Only those who responded to the questionnaire item about medical history in the health checkup survey were included. Individuals with a history of CVD were excluded from this study. The index date was set as the date of completing the health checkup questionnaire at a health checkup visit for the first time. CVDs that occurred within 60 days of the index date were considered as cases in which patients were diagnosed during the health checkup, and these CVD events were excluded as CVD outcomes. Thus, the outcome was defined as CVD events that occurred 60 days after the index date, and follow-ups ended on the date that CVD events occurred (ie, within 5 years from the index date), the date of the final hospital visit, or the date of death. As such, the time-at-risk period was set as 61 days after the index date to 5 years after the index date.

The primary outcome was the comparison of CVD hazard ratios (HRs) between the group that underwent CCTA (target group) and the group that did not undergo CCTA at the health checkup visit (comparison group).

In the subgroup analyses, the CVD HRs, which were based on the ACC/AHA ASCVD risk score and the FRS, were analyzed. The patients were stratified into the nonrisk and low-risk group

or the high-risk group based on a cutoff score of 10 for the FRS [7] and 5 for the ASCVD risk score [8].

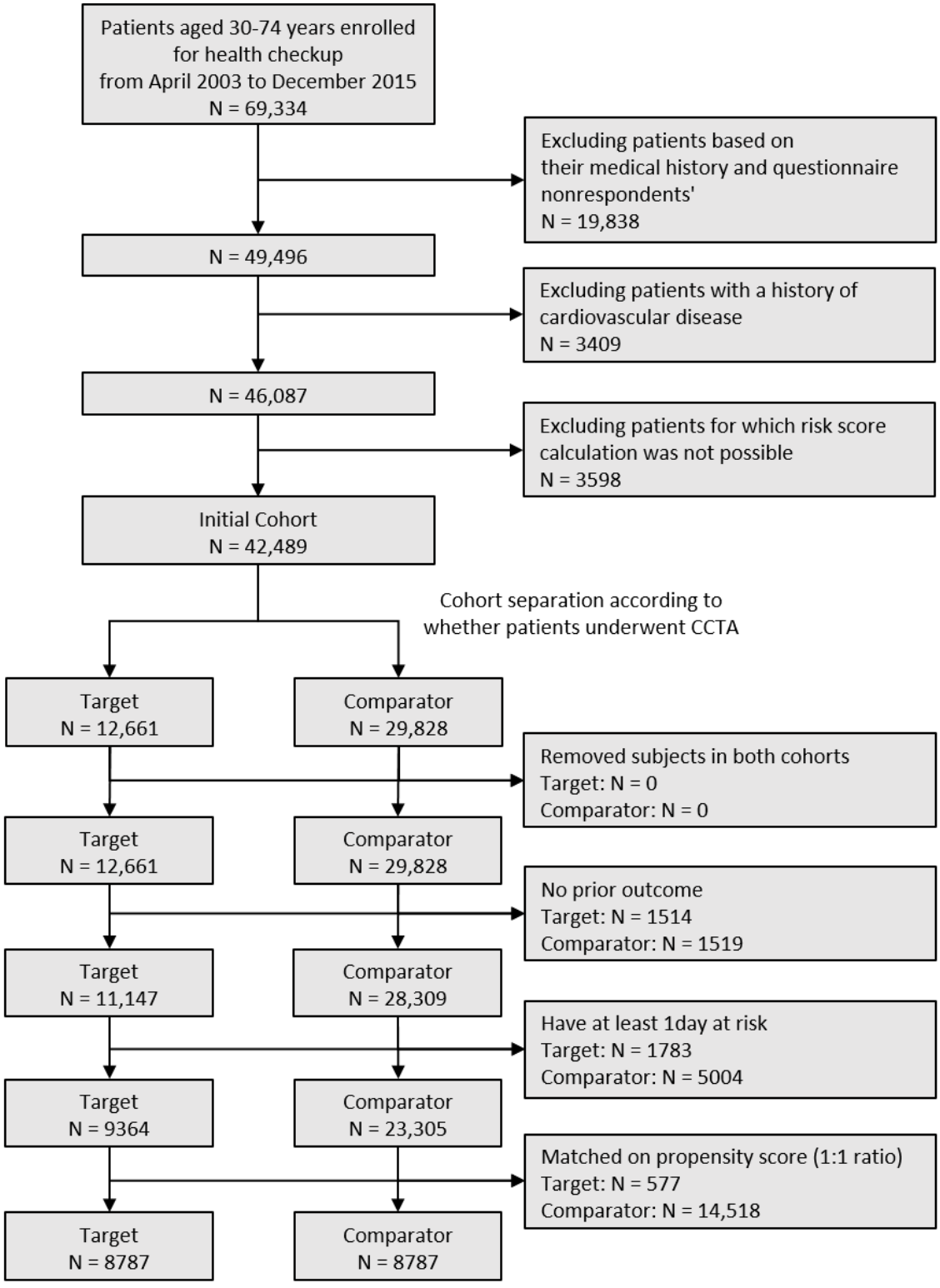
The secondary outcome was the difference between the risk scores of patients who underwent health checkups 2 years and 5 years after the index date. The differences between the risk scores at the index date and those at the times of subsequent examinations were used for comparative analyses.

Study Population

From April 2003 to December 2015, a total of 69,334 patients aged 30 to 74 years were enrolled for a health checkup. Of these patients, only 49,496 responded to the questionnaire, and only 46,087 patients had no cardiovascular history. A total of 42,489 patients for whom we could calculate the risk score—a key indicator of this study—were selected as the initial cohort.

Initially, of the 42,489 patients who were included in the analysis, 12,661 underwent CCTA (target group), and 29,828 did not (comparison group). Of these patients, 1514 from the target group and 1519 from the comparison group with a history of CVD before the index date were excluded from the analysis. In addition, 1783 patients from the target group and 5004 patients from the comparison group who did not fulfill the minimum observation period of 1 day during the time-at-risk window were excluded. The remaining 9364 patients from the target group and 23,305 patients from the comparator group underwent 1:1 propensity score matching. During 1:1 propensity score matching, 577 people who did not match the comparator group were excluded from the target group because matching was performed to maximize the minority group, and 14,518 people were excluded from the comparator group. Finally, 8787 of the 12,661 patients (69.4%) from the initial target cohort were selected as the final target group, and 8787 of the 29,828 patients (29.5%) from the initial comparator cohort were used for the analysis as the final comparator group (Figure 1).

Figure 1. The flowchart of the study population. CCTA: coronary artery computed tomography angiography.



Covariates

Approximately 13,000 variables were used as covariates for propensity score matching. These covariates included patient clinical data that were obtained at any time prior to the index date and health checkup data that were obtained on the index date. The patient clinical covariates included the condition era, the condition group era, the drug group era, observations, measurements, procedures, the Charlson Comorbidity Index score, the Diabetes Complications Severity Index score, the $CHADS_2$ (Congestive Heart Failure, Hypertension, Age, Diabetes, Previous Stroke/Transient Ischemic Attack [2 points]) score, the CHA_2DS_2-VASc (Congestive Heart Failure, Hypertension, Age≥75 [Doubled], Diabetes, Stroke [Doubled], Vascular Disease, Age 65 to 74, and Sex Category [Female]) score, and the hospital frailty risk score. The covariates that were measured at the index date included demographic data, such as sex, age, education level, average monthly income, and

marital status; health questionnaire data, such as any history of cancer and chronic diseases (hypertension, diabetes, and hyperlipidemia), medication history (antihypertensive drugs, antidiabetic drugs, antihyperlipidemic drugs, and aspirin), smoking status, and family history; and health checkup data, such as height, weight, BMI, blood pressure (systolic and diastolic), waist circumference, glucose levels, uric acid levels, aspartate aminotransferase levels, alanine aminotransferase levels, triglyceride levels, total cholesterol, high-density lipoprotein cholesterol levels, low-density lipoprotein cholesterol levels, and glycated hemoglobin A1c levels.

Outcomes

The outcome of this study was the first registered CVD event, which was based on a CVD diagnosis during the observation period. A CVD event was defined based on *International Classification of Diseases, 10th Revision* (ICD-10) codes I20 to I25 (ischemic heart disease), I50 (heart failure), I60 to I69 and G45 to G46 (stroke), and E78 (hypercholesterolemia). As we intended to assess the HRs of CVDs resulting from arteriosclerotic diseases only, we excluded cardiogenic diseases, such as atrial fibrillation and aneurysm (I42-I43, I48, I71, I62, and I68), and diseases caused by external accidental factors (I60 and I62). The ICD-10 codes that were chosen as the outcomes were reviewed by 1 clinical specialist and 1 nurse.

Statistical Analysis

We used the population-level estimation methodology and an open-source tool provided by OHDSI [9]. All analyses were performed by using R version 4.0.3 (R Foundation for Statistical Computing) [10]. Large-scale propensity score matching [11] was performed to adjust for potential confounding and to resolve the imbalance between the target and comparison cohorts caused by selection bias—a result of the retrospective observational nature of this study. The propensity score–matched model, which used approximately 13,000 covariates, was fitted through regularized regression, and the propensity score was calculated as the probability of a patient undergoing CCTA based on the covariates. Target and comparison group patients with similar propensity scores were matched to create a balanced cohort. To establish a matched cohort, we performed 1:1 propensity score matching by using a caliper width of 0.2 of the SD of the logit. The conditional Cox proportional hazards model was used to estimate HRs for the target group, in relation to the comparison group. The balance of the covariates between the cohorts was assessed based on the standardized difference of the mean (<0.1). Statistical significance was evaluated at $P<.05$ for 2-tailed tests.

To explain any residual bias after controlling for the measured covariates, we used negative control outcomes that were unlikely to be induced or prevented by undergoing CCTA; thus, the actual HR was anticipated to be 1. The negative control outcomes were selected by a clinical specialist through a manual review of the outcomes that were used in a previous OHDSI study [12] (Table S1 in Multimedia Appendix 1). The same study design was used to estimate the outcomes of interest and calculate the HR estimate for the negative control group, and all HR estimates were presented with 95% CIs and *P* values, along with the empirical null distribution and adjustment [13,14]. The empirical equivalence of the two cohorts was assessed by using the propensity score distribution. We also reported the power analysis; propensity score; cohort balance before and after propensity score matching; fitted null distribution; calibration chart for negative control outcomes; and Kaplan-Meier curve, which shows the proportional hazards assumption over time.

To confirm the changes in the differences in ASCVD risk scores and FRSs, we used the 2-group comparison method. The normality of the amount of change was confirmed by using the Shapiro-Wilk test, and the changes in the two groups were confirmed by using the Wilcoxon rank-sum test.

Results

Characteristics of Study Participants

Table 1 shows the baseline characteristics of the patients before and after propensity score matching. The table shows the patients' age groups, sex, and BMIs; the number of patients in the risk score groups; and the follow-up periods. For most demographic characteristics, the differences between groups decreased after matching. The standardized difference of the mean for the covariates decreased from 0.4 to 0.07 after propensity score matching, which is lower than the conventional standard of 0.1, thereby confirming that propensity score matching was performed correctly (Figure 2). This can also be observed in Figure 3, which compares the distributions from before and after propensity score matching.

Table 1. The baseline characteristics of the study population before and after propensity score matching.

Characteristics	Before matching			After matching		
	CCTA[a] group (n=12,661)	Non-CCTA group (n=29,828)	Standard difference	CCTA group (n=8787)	Non-CCTA group (n=8787)	Standard difference
Age group[b] (years), n (%)						
30-34	226 (1.8)	2442 (8.2)	−0.26	150 (1.7)	155 (1.8)	0
35-39	1043 (8.2)	4319 (14.5)	−0.19	761 (8.7)	700 (8)	0.03
40-44	1870 (14.8)	5257 (17.6)	−0.08	1406 (16)	1263 (14.4)	0.05
45-49	2516 (19.9)	5134 (17.2)	0.07	1846 (21)	1697 (19.3)	0.04
50-54	2435 (19.2)	4617 (15.5)	0.10	1702 (19.4)	1678 (19.1)	0.01
55-59	2084 (16.5)	3322 (11.1)	0.16	1373 (15.6)	1438 (16.4)	−0.02
60-64	1468 (11.6)	2275 (7.6)	0.14	908 (10.3)	1050 (11.9)	−0.05
65-69	734 (5.8)	1564 (5.2)	0.02	471 (5.4)	568 (6.5)	−0.05
70-74	285 (2.3)	898 (3.0)	−0.05	170 (1.9)	238 (2.7)	−0.05
Sex[b], n (%)						
Female	4757 (37.6)	12,650 (42.4)	−0.10	3561 (40.5)	3368 (38.3)	0.04
Male	7904 (62.4)	17,178 (57.6)	0.10	5226 (59.5)	5419 (61.7)	−0.04
BMI[b] (kg/m^2), mean (SD)	24.2 (3.1)	23.7 (0.2)	0.18	24.0 (3.1)	24.1 (3.1)	−0.03
ACC/AHA[c] ASCVD[d] risk score[e], n (%)						
High (≥5)	5036 (39.8)	8576 (28.8)	N/A[f]	3062 (34.8)	3493 (39.8)	N/A
Low (<5)	7625 (60.2)	21,252 (71.2)	N/A	5725 (65.2)	5294 (60.2)	N/A
Framingham risk score[e], n (%)						
High (≥10)	4996 (39.5)	8155 (27.3)	N/A	3030 (34.5)	3381 (38.5)	N/A
Low (<10)	7665 (60.5)	21,673 (72.7)	N/A	5757 (65.5)	5406 (61.5)	N/A
Follow-up period (days)[e], mean (SD)	2220.3 (1731.6)	1928.9 (1675.5)	N/A	2604 (1594.4)	2583.1 (1657.0)	N/A

[a]CCTA: coronary artery computed tomography angiography.

[b]Variables used in propensity score matching.

[c]ACC/AHA: American College of Cardiology/American Heart Association.

[d]ASCVD: atherosclerotic cardiovascular disease.

[e]Variables not used in propensity score matching.

[f]N/A: not applicable.

Figure 2. Standardized difference of means between the two groups of covariates before and after propensity score matching.

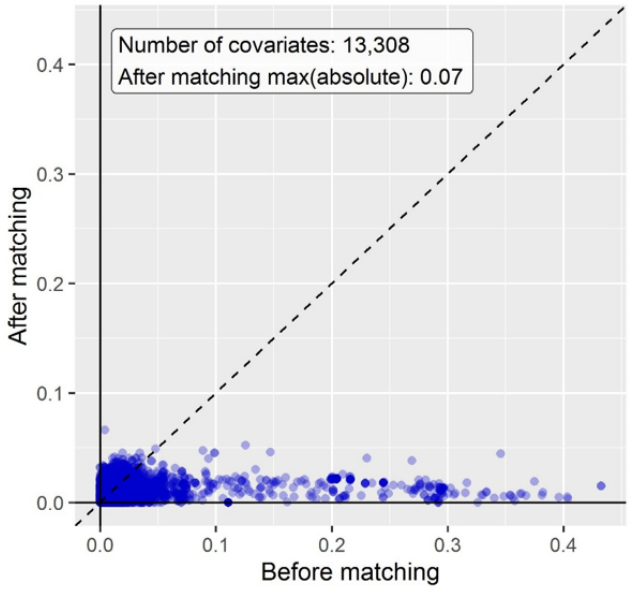
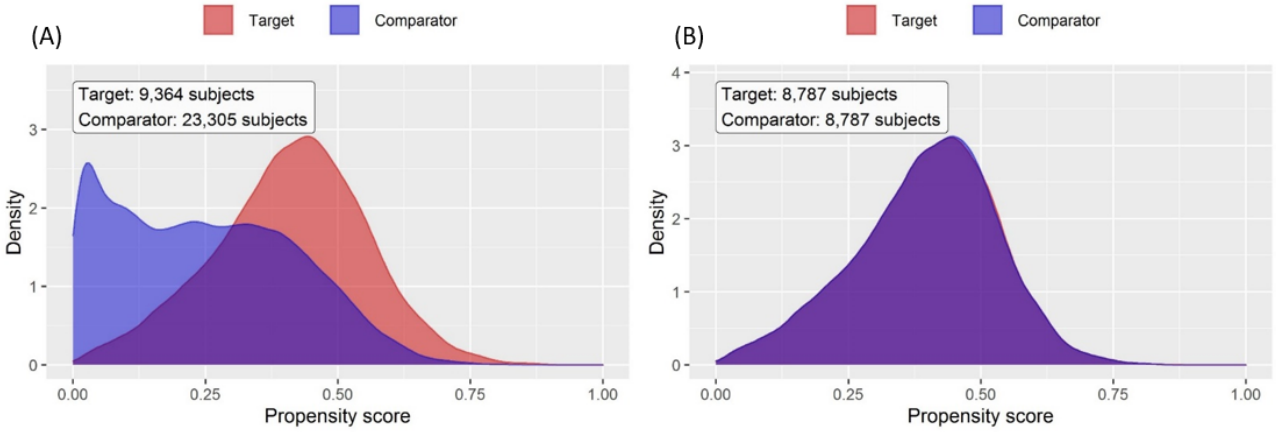


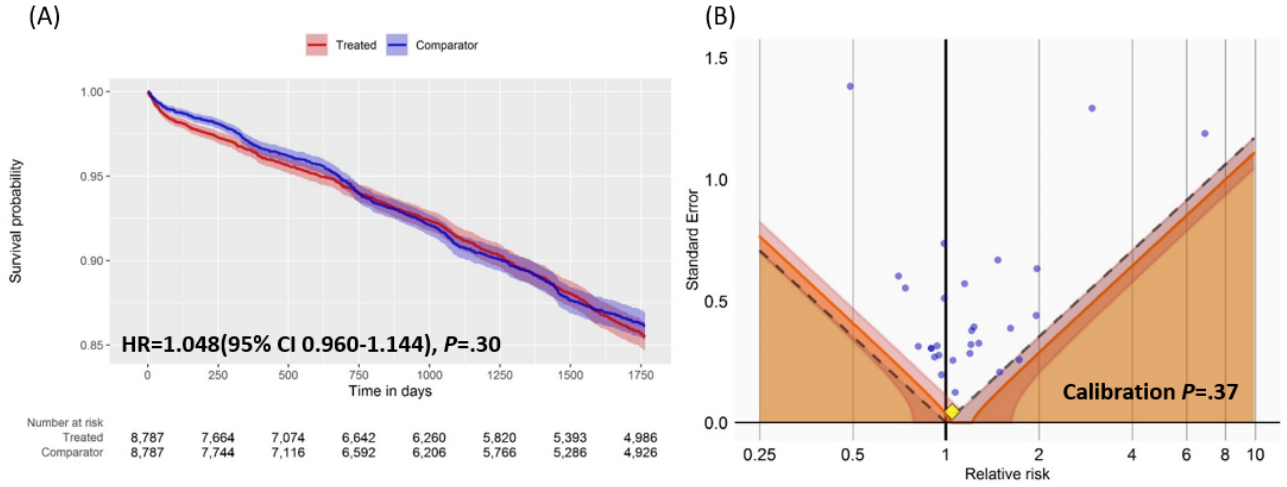
Figure 3. Distribution of propensity scores in each group (A) before and (B) after propensity score matching.



Effect of CCTA on CVDs

The Cox proportional hazards model was used to estimate and compare the HRs of CVDs among the target and comparison groups after propensity score matching, and no statistically significant differences were found between the two groups. The Kaplan-Meier analysis revealed that the HR was 1.048 (95% CI 0.960-1.144), which was not statistically significant (*P*=.30). The calibration *P* value, which was adjusted by using a negative control and was the most important indicator in our analysis, was .37, indicating no statistical significance (Figure 4).

Figure 4. (A) Kaplan-Meier curve plot and (B) rejection area plot with negative outcome controls applied in the main analysis. HR: hazard ratio.



Subgroup Analysis

The study population was stratified based on the cutoff scores for the ACC/AHA ASCVD risk score and FRS for subgroup analyses. Table 2 presents the results of each analysis. In each subgroup, the standardized difference of the mean dropped to <0.1 after propensity score matching. Figure S1 in Multimedia Appendix 1 shows the propensity score distributions, and Figure S2 in Multimedia Appendix 1 shows the standardized difference of the mean among groups of covariates before and after propensity score matching.

In the ASCVD high-risk subgroup (risk score≥5), 3149 patients were included in both the target group and comparison group. In the low-risk subgroup (risk score<5), 5524 patients were included in both the target group and comparison group. In the high-risk and low-risk subgroups, the calibration *P* value, which was adjusted by using negative controls, was .39 and .50, respectively, showing no significant differences in the HRs of CVDs among the target and comparison groups.

In the FRS high-risk subgroup (FRS≥10), 3110 participants were included in both the target group and comparison group. In the low-risk subgroup (FRS<10), 5602 patients were included in both the target group and comparison group. The calibration *P* value, which was adjusted by using negative controls, was .13 and .57 in the high-risk and low-risk subgroups, respectively, indicating no significant differences in the HRs of CVDs among the target and comparison groups (Figure S3 in Multimedia Appendix 1).

Table 2. The risk of cardio-cerebrovascular disease at 5 years in each subgroup based on the American College of Cardiology/American Heart Association (ACC/AHA) atherosclerotic cardiovascular disease (ASCVD) risk score and Framingham risk score (FRS).

	Hazard ratio (95% CI)	*P* value[a]	Calibration *P* value[b]
ACC/AHA ASCVD risk score			
High (≥5)	1.113 (0.984-1.259)	.09	.39
Low (<5)	0.999 (0.881-1.133)	.99	.50
FRS			
High (≥10)	1.166 (1.031-1.321)	.02	.13
Low (<10)	1.004 (0.883-1.141)	.96	.57

[a]Kaplan-Meier analysis *P* value.

[b]Calibration *P* value that was adjusted by using a negative control.

Risk Scores at 2 and 5 Years

The 2-year median change in the ASCVD risk scores and the FRSs of the non-CCTA group was 0.23 and 0.60, respectively. In contrast, the ASCVD risk scores and the FRSs of the CCTA group changed by 0.17 and 0.39, respectively. There was a statistically significant difference for both risk scores, with *P* values of .03 and .02, respectively.

The 5-year median change in the ASCVD risk scores and the FRSs of the non-CCTA group was 1.06 and 1.61, respectively. In contrast, the ASCVD risk scores and the FRSs of the CCTA group changed by 1.10 and 1.66, respectively. There was no statistically significant difference for both risk scores, with *P* values of .97 and .85, respectively (Table 3).

Table 3. Changes in the differences in American College of Cardiology/American Heart Association (ACC/AHA) atherosclerotic cardiovascular disease (ASCVD) risk scores and Framingham risk score (FRSs) from baseline at 2 and 5 years.

	CCTA[a] group		Non-CCTA group		*P* value[b]
	Patients, n	Change in score, median (IQR)	Patients, n	Change in score, median (IQR)	
Differences in risk scores from baseline at 2 years					
ACC/AHA ASCVD risk scores	1330	0.17 (−0.16 to 1.08)	1691	0.23 (−0.10 to 1.30)	.03
FRSs	1330	0.39 (−0.80 to 1.96)	1691	0.60 (−0.69 to 2.26)	.02
Differences in risk scores from baseline at 5 years					
ACC/AHA ASCVD risk scores	1232	1.10 (0.08 to 1.57)	1372	1.06 (0 to 2.79)	.97
FRSs	1232	1.66 (0.04 to 3.92)	1372	1.61 (0.09 to 4.11)	.85

[a]CCTA: coronary artery computed tomography angiography.

[b]Wilcoxon rank-sum test *P* value.

Discussion

Principal Results

From our population-level estimation study, which compared the CVD HRs of a health checkup group that was undergoing CCTA with those of a group that was not undergoing CCTA over 5 years, although some benefits were observed at 2 years, we found no significant difference (calibration *P*=.37) in the final risk of CVD events between the two groups. It seems that CCTA has no beneficial effect on CVD prevention for long periods of time.

Communication about medical examinations and examination results through counseling has been reported to improve health indicators, such as CVD risk. In the Korean national health insurance service screening program, the group that underwent cardiovascular health screening for 40-year-olds had higher rates of new hypertension, diabetes, and hyperlipidemia, whereas the incidence of CVD mortality, all-cause mortality, and major adverse cardiovascular events was lower [15]. Per the results of an analysis of the same data, the group that received counseling after the health checkup had higher motivation stages of health behavior change than those of the group that received only the checkup [16]. The smoking cessation rate was higher after 2 years when compared to that of the group who received only the checkup [17]. Engberg et al [18] reported that cardiovascular risk scores, BMIs, and serum cholesterol levels were lower in the intervention groups than those in the control group after 5 years' worth of health screenings and consultations.

In existing studies that require lifestyle modifications, such as modifications for obesity, smoking cessation, and substance abuse, the effects of 1-time interventions or short-term interventions, interviews, and counseling tend to weaken over time. In a study that used the motivational interview technique for people with substance abuse issues, the positive effect observed at 3 months disappeared at 12 months [19], and in another study, the effect of smoking cessation treatment continued for 10 weeks and gradually slowed down at 3, 6, and 12 months [20].

Our study compared patients who did or did not undergo additional coronary computed tomography. Both groups underwent the same levels of examination and counseling, which were conducted by the cardiovascular health screening program of the national service in 1 hospital.

Smoking status, blood pressure, and blood lipid concentration, which are major factors in the FRS and ASCVD pooled cohort equations score, are closely related to lifestyle changes. Similar to previous studies, the effect of a single coronary computed tomography scan and the results of counseling decreased over time, and the differences that were observed after 2 years disappeared after 5 years.

Limitations

This study has some limitations. First, the follow-up period was 5 years, and the risk scores were not observed for a longer period (eg, 10 years, as CVDs can last for >10 years). A follow-up study for identifying a risk score that is suitable for CVD prediction over longer periods can be conducted in the future. Second, as this was a single-center study, some of the outcomes may not be generalizable. Multicenter studies that use OHDSI data networks can provide more generalizable evidence. Third, this study included patients who visited the health promotion center multiple times; those who did not undergo CCTA at the first visit but underwent CCTA during subsequent visits were included in the comparison group. Therefore, the differences between the groups might have been attenuated. This can be avoided by conducting a prospective cohort study. Lastly, observational research that uses EHR data has the limitation that it cannot fully capture the entirety of a patient's health information [21]. This study converted EHR data into common data model data, and it has the same limitation. If the participants of this study underwent examinations and treatments outside of the hospital, there was a disadvantage that the records for these procedures were not recorded in the database. Additionally, with regard to drugs, the SNUBH common data model converted data on prescription drugs for outpatients and administration drugs for inpatients. Thus, it was not known whether the drugs ordered for the outpatients were taken on time by the patients. As such, selection bias may have occurred due to information not being recorded in the database. Although it is possible to reduce channeling bias through large-scale propensity score matching, which we used in this study, there may still be the limitation that such matching cannot reduce selection bias [22].

Comparison With Prior Work

Waugh et al [23] conducted a systematic review and meta-analysis of 5 studies and reported that computed tomography has no benefits as a screening tool for the potential onset of CVDs. However, a closer review revealed that all 5 included studies were inappropriate in terms of their findings about the prophylactic benefits of CCTA. All of these studies investigated the association between coronary artery calcium (CAC) and the onset of CVDs or death after a specific follow-up period in patients who underwent CCTA screening. They used a short follow-up period and analyzed the results in the context of the presence of CAC as opposed to CCTA findings. Therefore, the conclusion of the meta-analysis by Waugh et al [23]—CCTA screening is not effective—was based on the finding that the risk of heart disease was not elevated in people undergoing a CAC assessment via CCTA, as opposed to an assessment of the prophylactic benefits of CCTA itself. Further, since the measurement of CAC is regarded as a reliable method for CVD risk assessment, a study claimed that CCTA should be introduced for the screening of asymptomatic individuals [24]. However, other studies claim that CCTA is cost-ineffective, although these admit that CAC, when observed via CCTA, is a better predictor of CVD than the FRS [25]. We supplemented these studies by comparing groups that underwent CCTA with those that did not undergo CCTA.

McEvoy et al [26] examined the differences in the incidence of coronary artery disease between CCTA and comparison groups after a fixed follow-up period. The authors matched the propensity scores of 1000 individuals who underwent CCTA for a health checkup with those of 1000 individuals who did not undergo CCTA (ie, the comparison group) and compared the incidence of coronary artery disease at the 90-day and 18-month follow-ups. The study reported that CCTA-based screening was significantly associated with an increased rate of invasive tests and medication use but was not associated with the incidence of coronary artery disease, concluding that CCTA is not recommended for screening purposes. However, the study was limited by the small number of cases and the short follow-up periods.

Our study presents reliable evidence about CCTA, which was obtained by performing large-scale propensity score matching and using EHR and health checkup questionnaire responses from OMOP-CDM data. We studied a large study sample over a longer study period than those used by previous studies. Although past studies used either 90-day follow-ups or 18-month follow-ups, we observed the patients from 60 days after the index date to 5 years after the index date to analyze the CVD HRs in relation to CCTA. Moreover, while previous studies had approximately 1000 patients in both the target group and comparison group, we included 8787 patients in each group. The data were also standardized, which enabled us to perform an efficient analysis across organizations and use the same analysis codes. Future studies can investigate the effects of CCTA and CVD in larger populations over long follow-up periods, in collaboration with organizations that convert health questionnaire data into the common data model format.

We also stratified the population into high-risk and low-risk groups based on the ASCVD risk score and FRS. Even in the high-risk group, CCTA screening did not have a significant effect (ASCVD risk score: calibration P=.39; FRS: calibration P=.13) on the prevention of CVD.

Based on the changes in risk scores, a significant difference was observed between the CCTA and comparison groups after 2 years (change in ASCVD risk scores: P=.03; change in FRSs: P=.02). However, this difference was not significant after 5 years (change in ASCVD risk score: P=.92; change in FRSs: P=.85). We speculate that patients are motivated to manage their risk score factors for a brief period immediately after the CCTA test; however, the significance decreases over long periods.

Conclusions

Through a retrospective cohort study that was conducted over a 5-year period, we found that CCTA had no significant preventive effect on future CVDs. We also demonstrated the potential of converting health checkup data into OMOP-CDM data and integrating such data into common data model–based EHR data for research targeting the health checkup population. Although we examined the outcomes of CVDs after CCTA, future studies could examine patients' health behaviors following CCTA. It is expected that the use of common data model data will be expanded to multicenter studies.

Acknowledgments

This work was supported by the Technology Innovation Program (grant 20004927 for "Upgrade of CDM based Distributed Biohealth Data Platform and Development of Verification Technology"), which is funded by the Ministry of Trade, Industry & Energy (Korea).

Data Availability

Common data model data are designed to support a distributed research network. Thus, access to the data is restricted on internal private networks, and the data are not publicly available.

Authors' Contributions

WKB designed this study. JC drafted the manuscript and performed the data analyses. SK, BK, HB, and WS reviewed the data extraction and study design. WKB, JC, and SY inspected and revised the manuscript. SY supervised this study. All authors have read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary material.
[DOCX File , 2366 KB - medinform_v10i10e41503_app1.docx]

References

1. World Health Organization. Global status report on noncommunicable diseases 2014. World Helath Organization. 2014. URL: https://apps.who.int/iris/bitstream/handle/10665/148114/9789241564854_eng.pdf;jsessionid=81554E9296D494A6A458F32FE22DA357?sequence=1 [accessed 2022-09-28]
2. Gal R, van Velzen SGM, Hooning MJ, Emaus MJ, van der Leij F, Gregorowitsch ML, et al. Identification of risk of cardiovascular disease by automatic quantification of coronary artery calcifications on radiotherapy planning CT scans in patients With breast cancer. JAMA Oncol 2021 Jul 01;7(7):1024-1032 [FREE Full text] [doi: 10.1001/jamaoncol.2021.1144] [Medline: 33956083]
3. Parikh R, Patel A, Lu B, Senapati A, Mahmarian J, Chang SM. Cardiac computed tomography for comprehensive coronary assessment: Beyond diagnosis of anatomic stenosis. Methodist Debakey Cardiovasc J 2020;16(2):77-85 [FREE Full text] [doi: 10.14797/mdcj-16-2-77] [Medline: 32670467]
4. Marano R, Rovere G, Savino G, Flammia FC, Carafa MRP, Steri L, et al. CCTA in the diagnosis of coronary artery disease. Radiol Med 2020 Nov;125(11):1102-1113. [doi: 10.1007/s11547-020-01283-y] [Medline: 32964325]
5. Emaus MJ, Išgum I, van Velzen SGM, van den Bongard HJGD, Gernaat SAM, Lessmann N, Bragatston study group. Bragatston study protocol: a multicentre cohort study on automated quantification of cardiovascular calcifications on radiotherapy planning CT scans for cardiovascular risk prediction in patients with breast cancer. BMJ Open 2019 Jul 27;9(7):e028752 [FREE Full text] [doi: 10.1136/bmjopen-2018-028752] [Medline: 31352417]
6. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc 2012;19(1):54-60 [FREE Full text] [doi: 10.1136/amiajnl-2011-000376] [Medline: 22037893]
7. Sohn C, Kim J, Bae W. The framingham risk score, diet, and inflammatory markers in Korean men with metabolic syndrome. Nutr Res Pract 2012 Jun;6(3):246-253 [FREE Full text] [doi: 10.4162/nrp.2012.6.3.246] [Medline: 22808350]
8. Arshi B, van den Berge JC, van Dijk B, Deckers JW, Ikram MA, Kavousi M. Implications of the ACC/AHA risk score for prediction of heart failure: the Rotterdam Study. BMC Med 2021 Feb 16;19(1):43 [FREE Full text] [doi: 10.1186/s12916-021-01916-7] [Medline: 33588853]
9. Schuemie M, Suchard M, Ryan P. New-user cohort method with large scale propensity and outcome models. CohortMethod. URL: https://ohdsi.github.io/CohortMethod/ [accessed 2022-09-28]
10. R: The R Project for statistical computing. R Foundation for Statistical Computing. URL: https://www.r-project.org/ [accessed 2022-09-28]
11. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983 Apr;70(1):41-55 [FREE Full text] [doi: 10.1093/biomet/70.1.41]
12. Suchard MA, Schuemie MJ, Krumholz HM, You SC, Chen R, Pratt N, et al. Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: a systematic, multinational, large-scale analysis. Lancet 2019 Nov 16;394(10211):1816-1826 [FREE Full text] [doi: 10.1016/S0140-6736(19)32317-7] [Medline: 31668726]
13. Schuemie MJ, Ryan PB, DuMouchel W, Suchard MA, Madigan D. Interpreting observational studies: why empirical calibration is needed to correct p-values. Stat Med 2014 Jan 30;33(2):209-218 [FREE Full text] [doi: 10.1002/sim.5925] [Medline: 23900808]
14. Schuemie MJ, Hripcsak G, Ryan PB, Madigan D, Suchard MA. Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data. Proc Natl Acad Sci U S A 2018 Mar 13;115(11):2571-2577 [FREE Full text] [doi: 10.1073/pnas.1708282114] [Medline: 29531023]
15. Lee H, Cho J, Shin DW, Lee SP, Hwang SS, Oh J, et al. Association of cardiovascular health screening with mortality, clinical outcomes, and health care cost: a nationwide cohort study. Prev Med 2015 Jan;70:19-25. [doi: 10.1016/j.ypmed.2014.11.007] [Medline: 25445334]
16. Son KY, Lee CM, Cho B, Lym YL, Oh SW, Chung W, et al. Effect of additional brief counselling after periodic health examination on motivation for health behavior change [corrected]. J Korean Med Sci 2012 Nov;27(11):1285-1291 [FREE Full text] [doi: 10.3346/jkms.2012.27.11.1285] [Medline: 23166407]
17. Son KY, Shin DW, Yang HK, Yun JM, Chun SH, Lee JK, et al. Effect of one-time brief additional counseling on periodic health examination for 40- and 66-year-olds: 2-Year follow up of 101 260 participants. Geriatr Gerontol Int 2018 Feb;18(2):329-337. [doi: 10.1111/ggi.13175] [Medline: 29044867]

18. Engberg M, Christensen B, Karlsmose B, Lous J, Lauritzen T. General health screenings to improve cardiovascular risk profiles: a randomized controlled trial in general practice with 5-year follow-up. J Fam Pract 2002 Jun;51(6):546-552. [Medline: 12100779]
19. McCambridge J, Strang J. Deterioration over time in effect of motivational interviewing in reducing drug consumption and related risk among young people. Addiction 2005 Apr;100(4):470-478. [doi: 10.1111/j.1360-0443.2005.01013.x] [Medline: 15784061]
20. Cropsey K, Eldridge G, Weaver M, Villalobos G, Stitzer M, Best A. Smoking cessation intervention for female prisoners: addressing an urgent public health need. Am J Public Health 2008 Oct;98(10):1894-1901. [doi: 10.2105/AJPH.2007.128207] [Medline: 18703440]
21. Hersh WR, Weiner MG, Embi PJ, Logan JR, Payne PRO, Bernstam EV, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. Med Care 2013 Aug;51(8 Suppl 3):S30-S37 [FREE Full text] [doi: 10.1097/MLR.0b013e31829b1dbd] [Medline: 23774517]
22. Weinstein RB, Ryan P, Berlin JA, Matcho A, Schuemie M, Swerdel J, et al. Channeling in the use of nonprescription paracetamol and ibuprofen in an electronic medical records database: Evidence and implications. Drug Saf 2017 Dec;40(12):1279-1292. [doi: 10.1007/s40264-017-0581-7] [Medline: 28780741]
23. Waugh N, Black C, Walker S, McIntyre L, Cummins E, Hillis G. The effectiveness and cost-effectiveness of computed tomography screening for coronary artery disease: systematic review. Health Technol Assess 2006 Oct;10(39):iii-iv, ix-x [FREE Full text] [doi: 10.3310/hta10390] [Medline: 17018228]
24. Kondos GT, Hoff JA, Sevrukov A, Daviglus ML, Garside DB, Devries SS, et al. Electron-beam tomography coronary artery calcium and cardiac events: a 37-month follow-up of 5635 initially asymptomatic low- to intermediate-risk adults. Circulation 2003 May 27;107(20):2571-2576. [doi: 10.1161/01.CIR.0000068341.61180.55] [Medline: 12743005]
25. Naghavi M, Maron DJ, Kloner RA, Berman DS, Budoff M, Superko HR, et al. Coronary artery calcium testing: A call for universal coverage. Prev Med Rep 2019 Sep 02;15:100879 [FREE Full text] [doi: 10.1016/j.pmedr.2019.100879] [Medline: 31193256]
26. McEvoy JW, Blaha MJ, Nasir K, Yoon YE, Choi EK, Cho IS, et al. Impact of coronary computed tomographic angiography results on patient and physician behavior in a low-risk population. Arch Intern Med 2011 Jul 25;171(14):1260-1268. [doi: 10.1001/archinternmed.2011.204] [Medline: 21606093]

Abbreviations

ACC/AHA: American College of Cardiology/American Heart Association
ASCVD: atherosclerotic cardiovascular disease
CAC: coronary artery calcium
CCTA: coronary artery computed tomography angiography
CHA2DS2-VASc: Congestive Heart Failure, Hypertension, Age≥75 (Doubled), Diabetes, Stroke (Doubled), Vascular Disease, Age 65 to 74, and Sex Category (Female)
CHADS2: Congestive Heart Failure, Hypertension, Age, Diabetes, Previous Stroke/Transient Ischemic Attack (2 points)
CVD: cardio-cerebrovascular disease
EHR: electronic health record
FRS: Framingham risk score
HR: hazard ratio
ICD-10: International Classification of Diseases, 10th Revision
IRB: Institutional Review Board
OHDSI: Observational Health Data Sciences and Informatics
OMOP-CDM: Observational Medical Outcomes Partnership Common Data Model
SNUBH: Seoul National University Bundang Hospital

Edited by C Lovis; submitted 28.07.22; peer-reviewed by S Chang, K Adapa; comments to author 18.08.22; revised version received 04.09.22; accepted 24.09.22; published 13.10.22.

Please cite as:
Bae WK, Cho J, Kim S, Kim B, Baek H, Song W, Yoo S
Coronary Artery Computed Tomography Angiography for Preventing Cardio-Cerebrovascular Disease: Observational Cohort Study Using the Observational Health Data Sciences and Informatics' Common Data Model
JMIR Med Inform 2022;10(10):e41503
URL: https://medinform.jmir.org/2022/10/e41503
doi:10.2196/41503
PMID:36227638

Original Paper

Standardized Description of the Feature Extraction Process to Transform Raw Data Into Meaningful Information for Enhancing Data Reuse: Consensus Study

Antoine Lamer[1,2,3], PhD; Mathilde Fruchart[1], MSc; Nicolas Paris[3], MSc; Benjamin Popoff[4], MD; Anaïs Payen[1], PharmD; Thibaut Balcaen[5], MD; William Gacquer[6], MSc; Guillaume Bouzillé[7], MD, PhD; Marc Cuggia[7], MD, Prof Dr; Matthieu Doutreligne[8,9], MSc; Emmanuel Chazard[1], MD, Prof Dr

[1]Univ. Lille, CHU Lille, ULR 2694 - METRICS: Évaluation des Technologies de santé et des Pratiques médicales, Lille, France

[2]Fédération régionale de recherche en psychiatrie et santé mentale (F2RSM Psy), Hauts-de-France, Saint-André-Lez-Lille, France

[3]InterHop, Rennes, France

[4]Department of Anaesthesiology and Critical Care, Rouen University Hospital, Rouen, France

[5]Medical Information Department, Amiens-Picardy University Hospital, Amiens, France

[6]Digital Services Department, Amiens-Picardy University Hospital, Amiens, France

[7]Institut national de la santé et de la recherche médicale (INSERM), LTSI-UMR 1099, Univ Rennes, CHU Rennes, Rennes, France

[8]Mission Data, Haute Autorité de Santé, Saint-Denis, France

[9]SoDa project team, National Institute for Research in Digital Science and Technology (INRIA), Saclay-Île de France, Gif-sur-Yvette, France

Corresponding Author:
Antoine Lamer, PhD
Univ. Lille
CHU Lille
ULR 2694 - METRICS: Évaluation des Technologies de santé et des Pratiques médicales
1 place de Verdun
Lille, 59000
France
Phone: 33 320626969
Email: antoine.lamer@univ-lille.fr

Abstract

Background: Despite the many opportunities data reuse offers, its implementation presents many difficulties, and raw data cannot be reused directly. Information is not always directly available in the source database and needs to be computed afterwards with raw data for defining an algorithm.

Objective: The main purpose of this article is to present a standardized description of the steps and transformations required during the feature extraction process when conducting retrospective observational studies. A secondary objective is to identify how the features could be stored in the schema of a data warehouse.

Methods: This study involved the following 3 main steps: (1) the collection of relevant study cases related to feature extraction and based on the automatic and secondary use of data; (2) the standardized description of raw data, steps, and transformations, which were common to the study cases; and (3) the identification of an appropriate table to store the features in the Observation Medical Outcomes Partnership (OMOP) common data model (CDM).

Results: We interviewed 10 researchers from 3 French university hospitals and a national institution, who were involved in 8 retrospective and observational studies. Based on these studies, 2 states (track and feature) and 2 transformations (track definition and track aggregation) emerged. “Track” is a time-dependent signal or period of interest, defined by a statistical unit, a value, and 2 milestones (a start event and an end event). “Feature” is time-independent high-level information with dimensionality identical to the statistical unit of the study, defined by a label and a value. The time dimension has become implicit in the value or name of the variable. We propose the 2 tables “TRACK” and “FEATURE” to store variables obtained in feature extraction and extend the OMOP CDM.

Conclusions: We propose a standardized description of the feature extraction process. The process combined the 2 steps of track definition and track aggregation. By dividing the feature extraction into these 2 steps, difficulty was managed during track

definition. The standardization of tracks requires great expertise with regard to the data, but allows the application of an infinite number of complex transformations. On the contrary, track aggregation is a very simple operation with a finite number of possibilities. A complete description of these steps could enhance the reproducibility of retrospective studies.

(JMIR Med Inform 2022;10(10):e38936) doi:10.2196/38936

KEYWORDS

feature extraction; data reuse; data warehouse; database; algorithm; Observation Medical Outcomes Partnership

Introduction

The increasing implementation of electronic health records over the last few decades has made a significant amount of clinical data available in electronic format [1,2]. Originally, electronic health records were designed to collect and deliver data for health care, administrative, or billing purposes. In addition to these initial uses, they also offer opportunities for data reuse defined as "nondirect care use of personal health information" [3]. Thus, data reuse provides possibilities for research, quality of care assessment, activity management, or public health management [4-10].

When conducting research, the traditional approach consists of prospectively and often manually collecting simple and specific data according to the question addressed by the research protocol, using a clinical report form [11]. These data correspond to inclusion criteria and variables, that is, outcomes (eg, the length of stay in hospital or survival), exposures (eg, the taking of a drug or a surgery procedure), and adjusting variables (eg, age, sex, and patient history). When performing a prospective study, these data are defined upstream and are then collected manually in routine practice with human expertise, one record at a time, and background is taken into account. If needed, third-party data sources can be queried or caregiver expertise can be sought. This approach is expensive and time-consuming, and it generally results in only a limited sample size for a single use [7,11]. However, the final data set consists of explicit information that does not need further computation.

In contrast, data reuse builds on data sources already available at a low cost and offers a large volume of data [7]. Despite the many opportunities data reuse offers, its implementation presents many difficulties, and primary data cannot be reused directly. First, data reuse encounters data quality problems that arise from the manner in which the data were entered or collected [12-16], and it requires a phase of data cleaning to deduplicate, filter, homogenize, or convert raw data [17,18]. Moreover, information is not always directly available in the source database and needs to be computed later from raw data when defining an algorithm [19-23]. This is generally called "data transformation" [24], "data aggregation" [25,26], or "feature extraction" [27]. Even if feature extraction often approximately answers the question, the process is not easy and brings methodological issues. Indeed, features are extracted from a static database (already saved and closed) for patients for whom the care event has already been completed years earlier and for a large number of records. All scenarios must be taken into consideration so as to avoid having to modify the extracted records individually and by hand before the analysis. The method of extraction may have substantial effects on the features generated [28].

Lastly, the heterogeneity of local data models and vocabularies complicates the pooling of data and the sharing of algorithms, tools, and results [29-33]. Initiatives have emerged to promote the reuse of data through "large-scale clinical data sharing and federation" and the implementation of common data models (CDMs) [34-38]. Observational Health Data Sciences and Informatics (OHDSI) is a community developed from the Observational Medical Outcomes Partnership (OMOP) [39-42]. The OMOP CDM is dedicated to observational studies, medical product safety surveillance, comparative effectiveness research, and patient-level predictive modeling. In this context, the OHDSI community shares methods and tools for the use of the OMOP CDM, which standardizes the structure and vocabulary of observational data. Around 2000 collaborators from 74 countries were involved in the OHDSI community in mid-2022 [43]. Analyses could be successfully applied on this model and be used at different data sites around the world [44,45].

Beside clinical data tables, which are appropriate for the storage of individual low-level records (ie, procedure_occurrence, condition_occurrence, and measurement), the OMOP CDM was extended with 5 tables to store derived elements [46]. In particular, the EPISODE table stores the abstracted episodes of care previously defined [47,48] and allows the extraction of chemotherapy episodes from drug records in order to compare anticancer treatment trajectories [49].

Feature extraction methods are poorly described when applied to compute secondary information from retrospective databases. They also lack an approach to store features in a persistent way in a data warehouse. The purpose of this article is to propose a standardized description of the steps and transformations that could help researchers to implement and document feature extraction, and improve the reproducibility of retrospective studies. It also includes identifying how features could be stored in the schema of a data warehouse implemented with the OMOP CDM.

Methods

Overview

This study involved the following 3 main steps: (1) the collection of relevant study cases that applied feature extraction and were based on the automatic and secondary use of data; (2) the standardized description of the feature extraction process, including the concepts, their characteristics, and the methods that were common to the study cases; and (3) the proposal of convenient tables to store features in the OMOP CDM.

Ethics Approval

This study did not require ethics approval as no personal data were collected and no interventions were implemented.

Collection of Study Cases

We were seeking examples of retrospective observational studies for which feature extraction operations had to be implemented. These studies did not need to be conducted for a specific field of research, during a defined time period, or using a particular data model. The prerequisite was to have transformed raw data into usable information and to be able to describe the process. We focused on studies performed with structured data and did not investigate feature extraction from unstructured data such as text, images, videos, or sound. We contacted researchers from 7 teams involved in data reuse in France between September 1, 2021, and December 31, 2021.

We conducted individual interviews and obtained handwritten notes. The researchers were asked to describe (1) the objective of the study, (2) the database they used (ie, claims or clinical database), (3) the nature of the data and the terminologies, (4) the difficulties they encountered when extracting information from raw data, (5) the features they had to extract to achieve the objectives of the study, (6) the use they made of the features in the study (ie, inclusion criteria, explanatory variables, or response variables), and (7) the steps that composed the feature extraction and the parameters that characterized the features.

The inclusion criteria define the characteristics that subjects must have to be included in a study. They usually include age, type and stage of a disease, and surgical procedure. The response variable is the target of a question in the study or experiment. It is usually survival, length of hospital stay, recovery, or complication of a disease. The explanatory variable is that variable whose changes might affect the response variable. It may be exposure to an event or to a treatment.

The studies were carried out on the following 2 types of databases: claims databases and hospital clinical databases. These 2 sources are relational databases with a tabular format. Each table contains only 1 entity (eg, patients, stays, and diagnoses), and each row corresponds to 1 record. The tables are linked together by the mechanism of foreign keys, allowing the identification of all the data of a patient or a stay, whatever the category. Most of the columns are structured data (ie, 1 type and 1 value per cell). These databases are usually queried using the SQL language. They can then be processed with programming languages, such as R and Python, to recalculate new essential information or to adapt the structure of the data to be able to analyze them more easily.

The claims databases were the French national hospital discharge database, referred to as *Programme de médicalisation des systèmes d'information* (PMSI) [50], and the French national claims database, referred to as *Système National des Données de Santé* (SNDS) [51]. These nationwide databases collect standardized discharge reports for all inpatient stays in French nonprofit or for-profit hospitals. They include individual-level data about the dates of admission and discharge, the hospital code number, the sector code and outcome (ie, discharge, hospital transfer, and death), social demographics (ie, gender, age, and place of residence), diagnoses, and medical procedures performed during the hospital stay. The diagnoses are coded according to the French version of the International Statistical Classification of Diseases and Related Health Problems, 10th Revision (ICD10). The medical procedures are documented according to the *Classification Commune des Actes Médicaux* (CCAM). In addition to these data, the SNDS database includes consumption of care outside the hospital (ie, pharmacy visits, general medical reimbursements, and nursing care). Prescribed medications are documented with the Anatomical Therapeutic Chemical (ATC) system, an international classification system, or with the *Code Identifiant de la Présentation* (CIP13).

The clinical databases were local hospital data warehouses collecting all information about laboratory results, medical procedures, diagnoses, and types of medical units and transfers between them. Two databases included the details of anesthesia procedures (ie, the steps of the surgical procedures, drug administrations, and signals recorded by the equipment in the operating room, eg, mean arterial pressure, heart rate, and tidal volume) [52]. In these databases, vocabularies are local terminologies developed by the software editor and updated by the physician during practice. They cover drugs, measurements, and steps of the surgical procedure. The last database was the Medical Information Mart for Intensive Care III database, a large open-source medical record database of critical care stays, publicly available in PhysioNet [53,54]. Diagnoses are documented with the International Statistical Classification of Diseases and Related Health Problems, 9th Revision (ICD9), and the procedures are documented with the Current Procedural Terminology.

Standardized Description of the Feature Extraction

In the second step, we performed a hierarchical analysis of the task (HAT) [55]. A HAT allows an understanding of the tasks that users need to accomplish in order to achieve certain goals. These tasks may be decomposed into several levels of subtasks, up to having atomic operations. In this study, we carried out a HAT to (1) understand the steps and transformations that the researchers had to implement to transform raw data into features and (2) identify the successive states of data, from raw data to features, describing the complexity and time dependency.

To do so, we asked them to describe the raw data they had at the beginning, and which were the different transformations they had to chain to obtain features. At each step, we described the complexity and time dependency. We have illustrated the succession of subtasks for each case study, in collaboration with the researcher involved in the study. From the obtained task descriptions and illustrations, we grouped the tasks according to the types of input and output data. Lastly, we propose a description of these different states and transformations, based on what was common to the study cases.

Evaluation of Feature Storage Possibilities in the OMOP CDM

In the last part, we studied the existing tables of the OMOP CDM that could allow the storage of features without losing information, that is, with adequate fields. In the reverse case, we would propose new tables to conform to the OMOP standard.

We would also define the attributes that would have to respect the OMOP standard and keep track of how features were computed to ensure the reproducibility of the studies.

Results

Collection of Study Cases

Among the 15 people we contacted, 3 did not answer and 2 reported not performing feature extraction. Based on the semistructured interviews, we collected 8 retrospective and observational studies from teams in 3 French university hospitals (Amiens, Lille, and Rouen) and the French high authority of health. Two of the studies were multisite studies, 4 used claims databases, and 5 used clinical databases.

The features identified represented different types of variables used for conducting retrospective analyses: inclusion criteria, explanatory variables, and response variables. Generic features were (1) occurrences of diagnoses, medical procedures, and age as inclusion criteria; (2) occurrences of medical procedures, occurrences of drug administrations, and transformations of vital signs as explanatory variables; and (3) hospital and intensive care mortality, hospital stay duration, and passage in intensive care as response variables. The study cases and the more complex features reported by the researchers are described in Table 1.

These various study cases were based on complex (ie, heterogeneous, multidimensional, unbalanced, and time-dependent) raw data. The heterogeneity of these raw data comes from the diversity of the variables involved to extract secondary computed features. The first 5 study cases (SC1-5) used measurements and transformed vital signs (arterial pressure and heart rate) or ventilatory signals (partial pressure of oxygen and tidal volume), SC6 and SC7 used drug administrations, and SC7 used laboratory results. In addition to their heterogeneity, the databases are multidimensional, which implies that the tables that compose them have different dimensions (ie, statistical units). Thus, each patient will have a different number of records in the other tables (procedures, diagnoses, measurements, drugs, etc), depending on the length of hospital stay, the care received, and the duration of follow-up. This number of different records from one patient to the other should however be reduced to one line per statistical unit of the study. Next, the modalities of variables are numerous and unbalanced, that is, each terminology has thousands of codes, some of which are widely used, while others are almost never needed. As a result, at the time of feature extraction, these thousands of codes generate as many columns, with, for example, features reporting the code as absent/present or the number, or reporting the number of times it has been documented. At last, raw data are time-dependent variables, that is, variables that are not necessarily constant over the course of the study.

Table 1. Description of study cases involving feature extraction for retrospective observational studies.

Study case	Objective of the study	Features needed to achieve the objectives of the study
SC1: Detection of hyperoxemia in mechanically ventilated patients	To evaluate the effect of hyperoxemia on ICU[a] mortality, during the first 24 h of ICU stay, in mechanically ventilated patients with septic shock according to the SEPSIS-3 criteria [56]	Explanatory variable: Weighted average of PaO_2[b] for mechanically ventilated patients with septic shock according to the SEPSIS-3 criteria. The measurements are recorded at irregular intervals. The signal is reconstructed to give one measurement per second.
SC2: Duration of hypotension during heavy surgery	To evaluate the impact of early blood pressure control in heavy surgeries on in-hospital mortality and length of stay	Explanatory variable: Duration of arterial pressure spent with a drop of 10% from the average value, during the procedure.
SC3: Duration of hypotension during cesarean section with spinal anesthesia	To characterize the effect of hypotension during cesarean section with spinal anesthesia on fetal pain	Explanatory variable: Duration of systolic arterial pressure with a drop of 20% from a reference value between induction and birth for a cesarean section with spinal anesthesia. The reference value is the mean value of the systolic arterial pressure between arrival in the operating room and the induction.
SC4: Heart rate and administration of atropine	To assess the evolution of heart rate before and after the administration of atropine (a medication used to treat bradycardia)	Explanatory variables: The median, minimum, and maximum values of heart rate are computed during 2 periods of 10 minutes, designed around the administration of atropine.
SC5: Compliance with ventilatory guidelines	To evaluate whether the recommendations in terms of ventilation in the operating room have been carried out [57]	Explanatory variable: End-tidal volume <8 mL/kg of ideal body weight during surgery.
SC6: Potentially inappropriate medications	To measure the impact of a therapeutic optimization intervention included in an integrated care pathway on PIM[c] prevalence and on hospital readmission in frail older people	Explanatory variable: Number of drug administrations from the French Laroche list [58] (potentially inappropriate medications) in the 90 days preceding the hospitalization. Number of drug administrations from the French Laroche list in the 90 days following the hospitalization.
SC7: Drug-drug interactions	To estimate the probability of the occurrence of INR[d] changes for each DDI[e] rule involving VKA[f] [59]	Explanatory variable: Administration of VKA with another drug defined in a DDI rule. Raw ATC[g] codes are mapped to wider categories by taking into account the active substances and the administration route. The period of interest started the day after the 2 drugs had been administered together and ended 4 days after the first of the 2 drugs was discontinued. Response variable: VKA potentiation with at least one value of INR ≥5 or VKA inhibition with at least one value of INR ≤1.5.
SC8: Compliance with guidelines for COPD[h] patients	To assess the percentage of suspect COPD patients having functional respiratory exploration for diagnosis	Explanatory variable: Suspect COPD patients defined as patients aged more than 40 years with one of several of the following treatments: bronchodilators, 3 antibiotic therapies for respiratory infection, or nicotinic substitutes.

[a]ICU: intensive care unit.

[b]PaO_2: partial pressure of oxygen.

[c]PIM: potentially inappropriate medication.

[d]INR: international normalized ratio.

[e]DDI: drug-drug interactions.

[f]VKA: vitamin K antagonist.

[g]ATC: Anatomical Therapeutic Chemical.

[h]COPD: chronic obstructive pulmonary disease.

Standardized Description of the States and Transformations Related to Feature Extraction

Figure 1 provides the complete description of SC6. First, raw records of administrative data were transformed into a new type of record corresponding to the occurrence of hospital stay (step 1). We will refer to this period as "track" in the rest of the manuscript. Then, this track was transformed to obtain a second track representing the 90 days before hospital stay (90_days) (step 2). Drug administrations included in the Laroche list were identified from raw records, and the periods of administration of drug A and drug B were computed based on the dates of

administration and the duration of treatment, in steps 3 and 4, respectively. Similar tracks were computed for all drugs included in the Laroche list, but for the clarity of the figure, we have chosen to illustrate only the first 2 drugs. After these 4 steps, comparisons between tracks were realized successively. This allowed comparisons of the tracks of administration of drug A and drug B to track 90_days, in steps 5 and 6, respectively. The results were joined in a common track to obtain the tracks of the administration of Laroche list items during track 90_days (step 7). Lastly, the number of distinct items was counted to obtain the final feature, that is, the number of drugs from the Laroche list administered in the 90 days preceding the hospital stay.

Table 2 summarizes these transformations, as well as the input and output data of each transformation. Standardized descriptions of all other study cases and feature extraction processes are available in Multimedia Appendix 1 and Multimedia Appendix 2.

Figure 1. Standardized description of study case 6.

Table 2. Input data, transformations, and output data for each step involved in the feature extraction of study case 6 (potentially inappropriate medications).

Step	Input data	Transformation	Output data
1	Raw data: Hospital stay	Selection of fields "admission date" and "discharge date"	Track: Hospital stay
2	Track: Hospital stay	Computing the previous 90 days	Track: 90 days before hospital stay
3	Raw data: Drug administration	Selection of drugs included in the Laroche list	Track: Drug A
4	Raw data: Drug administration	Selection of drugs included in the Laroche list	Track: Drug B
5	Track: 90 days before hospital stay + Track: Drug A	Intersection of the 2 tracks	Track: Drug A (Laroche)/90 days before hospital stay
6	Track: 90 days before hospital stay + Track: Drug B	Intersection of the 2 tracks	Track: Drug B (Laroche)/90 days before hospital stay
7	Track: Drug A (Laroche)/90 days before hospital stay + Track: Drug B (Laroche)/90 days before hospital stay	Union of the 2 tracks	Track: Drug Laroche list/90 days before hospital stay
8	Track: Laroche list/90 days before hospital stay	Count distinct (drug Laroche list/90 days before hospital stay)	Feature: Number of drugs from the Laroche list prescribed in the 90 days before hospital stay

States and Transformations

Based on the study cases and the HAT, we identified that data went through 2 states (track and feature) and benefited from 2 transformations (track definition and track aggregation). Table 3 summarizes the differences between the raw data, track, and feature, as well as the definitions of the 2 transformations. The whole process of feature extraction is illustrated for several types of raw data in Figure 2, and is fully described below.

The step of *track definition* aims at reducing the dimensions of raw data to the statistical unit of the study, which is the element of the population on which the statistical study is conducted. The statistical unit may refer to not only a patient, but also a hospital, hospital stay (SC6), specialized unit stay (SC1), or a procedure (SC2, SC3, SC4, and SC5), depending on the purpose of the study. During track definition, the data may be rebuilt or computed based on operations such as the selection of variables and values, the mappings between codes of terminologies (SC6 and SC7), the detection of the passage of values beyond a threshold (SC2 and SC3), or the application of any other expert rule (SC5, SC6, and SC7).

Track is an intermediate state between raw data and features. It results from the first operation and remains a time-dependent signal, defined by a statistical unit, a type of track, a value, or a set of values. The type of track may be the passage in a care unit, the administration of a drug, a health condition characterized by a diagnosis, or a heart rate signal. The value represents the track state, with a binary value for an on/off state or a quantitative value for a signal. Conditional operations may also be applied between tracks to generate new ones (eg, for detecting the simultaneous administration of 2 drugs). Based on this definition, Table 4 presents the tracks for the 8 study cases.

The step of *track aggregation* extracts final information from tracks during a specified period of interest. The extraction method reduces the multidimensionality and releases from the dependence on time. These methods are usual statistical functions (eg, minimum, maximum, mean, median, count, duration, and delay).

The *period of interest* is defined by a start date and an end date, which may come from different sources as follows: the administration of a drug, the step of a procedure, the visit with a health care professional, or the visit to a health care unit. For each date, there could be more than one candidate event. For example, in SC3, the start of the anesthesia procedure may be documented with 4 different events as follows: induction event, hypnotic administration, intubation, and mechanical ventilation. In the same way, the end of the anesthesia procedure may be defined by the following 2 events: extubation or the end of the anesthesia event. In this case, a priority rule based on expert knowledge or an aggregation operation (first or last event) selects the main event. Lastly, a time interval may be added to the start and end dates of the period to create an artificial period as follows: the 90 days preceding or following hospitalization (SC6).

At the end of the process, *feature* is a single value associated with a label (the feature name). In a feature, time is implicit and is no longer formalized by a date in the record. It may be sometimes represented in the name of the variable, with, for example, the mean value of arterial pressure before induction (eg, mean_map_before_induction). It may also be represented in the value of the feature itself (eg, for a delay or a duration). The feature depends greatly on the context of the study; thus, in SC2 and SC3, the same raw signal produces 2 distinct features that depend on the extraction methods and the periods of interest. Table 5 describes the features identified in our 8 study cases, according to the statistical unit, period, signal, and extraction method.

Table 3. Definitions and comparisons of the states and transformations involved in the feature extraction.

States and transformations	Description	Example	Time dimension	Complexity
Raw data (state)	Heterogeneous, multidimensional, and time-dependent low-level clinical data: demographic data, patient flow, laboratory results, drug administrations, procedures, diagnoses, and measurements. The time dimension is always beside the value as an attribute.	Raw measurements of mean arterial pressure	Yes	Yes
Track definition (transformation)	Reduction of the initial dimensions to the statistical unit and standardization of the data representation through an infinite possibility of operations with high expert knowledge. Conservation of the time dimension. Conditional operations may be performed on tracks to generate new tracks.	Resampling of the signal	Yes	Reduced
Track (state)	Homogeneous and time-dependent signal, defined by a homogeneous statistical unit, a type of track, and a set of time-stamped values. The time dimension remains beside each track.	Resampled signal with one measurement per second	Yes	No
Track aggregation (transformation)	Reduction of the time dimension: a period of interest, a track, and an extraction method based on a finished number of operations (minimum, maximum, median, sum, count, etc). The time dimension is reduced to obtain a single value, with time embedded in the variable name or inside the value.	Aggregation (minimum and mean values) of measurements recorded between the start and end of the anesthesia procedure	Reduced	No
Feature (state)	Time-independent high-level information with dimensionality identical to the statistical unit of the study, defined by a label and a value. The time dimension has become implicit in the value (eg, in a delay or a duration) or name of the variable (eg, a value at day 1).	Minimum and mean values of mean arterial pressure during the anesthesia procedure	Implicit	No

Figure 2. Feature extraction process transforming raw data into features.

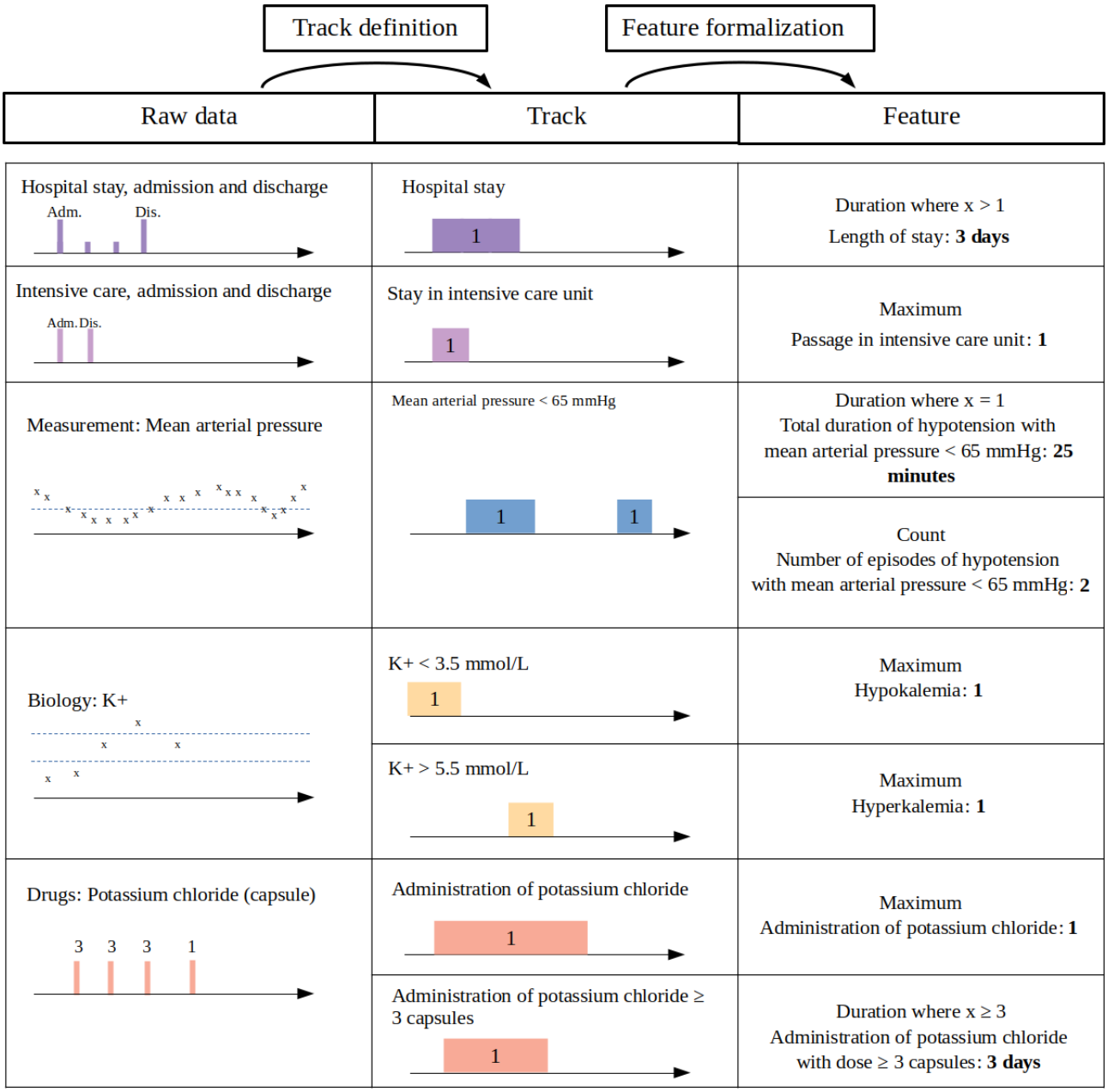


Table 4. Definition of tracks used in the study cases.

Study case and statistical unit	Track	Value(s)
SC1: Hyperoxemia in mechanically ventilated patients		
ICU[a] stay	First 24 hours of ICU stay for mechanically ventilated patients with septic shock	ICU stay=1
ICU stay	Resampled PaO_2[b]	PaO_2 repeated measurements
SC2: Duration of hypotension during general anesthesia		
Heavy surgery	General anesthesia procedure	General anesthesia procedure=1
Heavy surgery	Average value of mean arterial pressure	Average value
Heavy surgery	Episode of mean arterial pressure below 90% of the average value	Episode=1
SC3: Duration of hypotension during cesarean section with spinal anesthesia		
Cesarean section with spinal anesthesia	Arrival in the operating room to induction of anesthesia	Reference period=1
Cesarean section with spinal anesthesia	Induction of anesthesia to birth	Spinal anesthesia=1
Cesarean section with spinal anesthesia	Average value of the systolic arterial pressure between arrival in the operating room and induction of anesthesia	Average value
Cesarean section with spinal anesthesia	Episode of systolic arterial pressure below 80% of the average value	Episode=1
SC4: Heart rate and administration of atropine		
Administration of atropine	Before administration of atropine	Before=1
Administration of atropine	After administration of atropine	After=1
SC5: Compliance with ventilatory guidelines		
Anesthesia procedure with mechanical ventilation	Surgery	Surgery=1
SC6: Potentially inappropriate medications		
Hospital stay	Before hospital stay	Before hospital stay=1
Hospital stay	After hospital stay	After hospital stay=1
Hospital stay	Administration of drug X from the Laroche list	Drug X=1
SC7: Drug-drug interactions		
Patient	Administration of drug X (raw code)	Drug X=1
Patient	Administration of a drug family (ATC[c] category)	ATC category=1
Patient	Concomitant administration of a VKA[d] with a drug defined in a DDI[e] rule	Concomitant administration=1
Patient	INR[f] ≥5	Episode of INR ≥5
Patient	INR ≤1.5	Episode of INR ≤1.5
Patient	Concomitant administration of a VKA with a drug defined in a DDI rule and INR ≥5	VKA potentiation=1
Patient	Concomitant administration of a VKA with a drug defined in a DDI rule and INR ≤1.5	VKA inhibition=1
SC8: Compliance with guidelines for COPD patients		
Patient	Administration of one of several drugs among bronchodilators or nicotinic substitutes (ATC codes)	Drug X ≥1
Patient	Administration of 3 antibiotic therapies for respiratory infection (ATC codes)	Drug X ≥3

Study case and statistical unit	Track	Value(s)
Patient	Exposure to at least one of the drugs specific to suspected COPD[g]	Exposure to COPD-specific drugs=1
Patient	Induction of spirometry or functional respiratory exploration	Episode=1

[a]ICU: intensive care unit.

[b]PaO_2: partial pressure of oxygen.

[c]ATC: Anatomical Therapeutic Chemical.

[d]VKA: vitamin K antagonist.

[e]DDI: drug-drug interaction.

[f]INR: international normalized ratio.

[g]COPD: chronic obstructive pulmonary disease.

Table 5. Definitions of the characteristics for each feature of the study cases.

Study case	Statistical unit	Period	Track	Extraction method
SC1: Hyperoxemia in mechanically ventilated patients	ICU[a] stay	First 24 hours of ICU stay for mechanically ventilated patients with septic shock	Resampled PaO_2[b]	Weighted average
SC2: Hypotension during anesthesia	General anesthesia procedure	Anesthesia period	Mean arterial pressure	Sum of the duration of episodes of mean arterial pressure with a drop of 10% from the reference value
SC3: Duration of hypotension during cesarean section with spinal anesthesia	Cesarean section with spinal anesthesia	Anesthesia period	Systolic arterial pressure	Total duration of systolic arterial pressure below 80% of the reference value
SC4 :Heart rate and administration of atropine	Administration of atropine	Periods of 10 minutes before and after the administration of atropine	Heart rate	Median, minimum, and maximum values of heart rate
SC5: Compliance with ventilatory guidelines	Anesthesia procedure with mechanical ventilation	Surgery period	End-tidal volume	Mean end-tidal/ideal body weight >8
SC6: Potentially inappropriate medications	Hospital visit	Before hospital stay; after hospital stay	Administration of medications	Count of inappropriate drug administration according to the French Laroche list.
SC7: Drug-drug interactions	Patient	Day after the 2 drugs have been administered together and until 4 days after the first of the 2 drugs was discontinued.	Concomitant administration of a VKA[c] with a drug defined in a DDI[d] rule and INR[e] ≥5. Concomitant administration of a VKA with a drug defined in a DDI rule and INR ≤1.5.	Count of VKA potentiation. Count of VKA inhibition.
SC8: Compliance with guidelines for COPD[f] patients	Patient	Year following exposure to one of the drugs specific to COPD	Administration of medications	Count of the administration of drugs specific to COPD Binary indicator of FRE[g] induction

[a]ICU: intensive care unit.

[b]PaO_2: partial pressure of oxygen.

[c]VKA: vitamin K antagonist.

[d]DDI: drug-drug interaction.

[e]INR: international normalized ratio.

[f]COPD: chronic obstructive pulmonary disease.

[g]FRE: functional respiratory exploration.

Evaluation of Feature Storage Possibilities in the OMOP CDM

Five tables already exist in the OMOP CDM (DRUG_ERA, DOSE_ERA, CONDITION_ERA, EPISODE, and EPISODE_EVENT) for storing elements derived from raw data [46]. These tables cover the storage of spans of time when the patient is exposed to a specific drug ingredient (DRUG_ERA), when the patient is exposed to a constant dose of a specific drug ingredient (DOSE_ERA), or when the patient is assumed to have a given condition (CONDITION_ERA). These existing tables are suitable for pharmacoepidemiology studies with the comparison of periods of drug exposure and the resulting adverse events or evolution of the disease. The studies require only diagnosis and medication data from the tables CONDITION_OCCURRENCE and DRUG_EXPOSURE [39].

However, other types of data also need to be retransformed to obtain usable information for statistical analysis (in particular, procedures, measurements, biology results, or any types of steps in patient care). At this point, 2 alternatives allow other types of derived elements to be stored. The first approach involves adding an era table for each raw information that can be transformed into an era (ie, a measurement era, procedure era, biology era, etc). The second approach involves proposing a generic era table that would cover all types of raw data. With these 2 approaches, there would still be a lack of storage for the final features, which do not have the same structure as eras or episodes, since they are only an association of a value and a label, independent of time.

For this reason, on the one hand, the table TRACK could complement the model and store intermediate data (ie, all types of tracks and eras), which would ultimately be used to compute features, and on the other hand, the table FEATURE could extend the OMOP CDM for storing secondary computed data from measurements, procedures, observations, and stays, which would be used for the analysis and would need to be stored on a long-term basis.

These 2 new conceptual tables are illustrated in Figure 3. They comply with the OMOP guidelines in terms of field name and table organization [60]. For both tables, foreign keys reference the person, the visit, the visit details, the main concept (TRACK_CONCEPT_ID and FEATURE_CONCEPT_ID), and the type of this concept (TRACK_TYPE_CONCEPT_ID and FEATURE_TYPE_CONCEPT_ID). Similarly, the 2 tables provide core fields to store continuous values (VALUE_AS_NUMBER) or categorical values (VALUE_AS_CONCEPT_ID). The specificity of TRACK involves the preservation of the time dimension through the fields TRACK_START_DATE and TRACK_END_DATE. In the FEATURE table, in the case where a patient could present the same feature several times (eg, on different days), a foreign key to the EPISODE table allows differentiation of the occurrences of a feature [47]. Both tables also have the usual fields to store the source values expressed with local vocabularies.

Figure 3. Data model for the storage of periods and features in a relational database, compliant with the Observational Medical Outcomes Partnership (OMOP) common data model. FK: foreign key; PK: primary key.

TRACK		
TRACK_ID	bigint	PK
PERSON_ID	bigint	FK
TRACK_CONCEPT_ID	bigint	FK
TRACK_START_DATE	date	
TRACK_START_DATETIME	datetime	
TRACK_END_DATE	date	
TRACK_END_DATETIME	datetime	
TRACK_PARENT_ID	bigint	
TRACK_OBJECT_CONCEPT_ID	integer	FK
TRACK_TYPE_CONCEPT_ID	integer	FK
VALUE_AS_NUMBER	float	
VALUE_AS_CONCEPT_ID	bigint	FK
VISIT_OCCURRENCE_ID	bigint	FK
VISIT_DETAIL_ID	bigint	FK
TRACK_SOURCE_VALUE	varchar(50)	
TRACK_SOURCE_CONCEPT_ID	bigint	FK

FEATURE		
FEATURE_ID	integer	PK
PERSON_ID	integer	FK
FEATURE_CONCEPT_ID	integer	FK
FEATURE_TYPE_CONCEPT_ID	integer	FK
PERIOD_ID	integer	FK
VALUE_AS_NUMBER	numeric	
VALUE_AS_CONCEPT_ID	integer	FK
UNIT_CONCEPT_ID	integer	FK
VISIT_OCCURRENCE_ID	integer	FK
VISIT_DETAIL_ID	integer	FK
FEATURE_SOURCE_CONCEPT_ID	integer	FK
FEATURE_SOURCE_VALUE	varchar(50)	
UNIT_SOURCE_VALUE	varchar(50)	
VALUE_SOURCE_CONCEPT_ID	integer	FK
VALUE_SOURCE_VALUE	varchar(50)	

Discussion

Principal Findings

In this article, we propose a standardized description of the feature extraction process, which is implemented when transforming heterogeneous, multidimensional, and time-dependent raw data into valuable information for conducting observational retrospective studies. The process combines 2 steps (track definition and track aggregation). Track definition aims at transforming raw data into multiple tracks representing the periods of interest or reconstructing a signal. Track aggregation computes usable information from a final track for applying an extraction method during a period of interest. The resulting features are the 1-dimensional and time-independent variables that will be included in the statistical analysis.

By dividing the feature extraction into these 2 steps, the difficulty is managed during track definition. The first step aims at creating tracks, with a common unit adequate for the statistical unit of the study and a homogeneous temporal scale. Tracks then allow the application of an infinite number of complex transformations, such as the mapping of concepts for the detection of drug-drug interactions (SC7). These transformations require great expertise with regard to the data and are mainly implemented on a custom basis. On the contrary, track aggregation is a very simple operation, with a finite number of possibilities.

Strengths of the Study

The definitions of the transformations are based on various cases, and they were carried out on different databases from several centers. Feature extraction is the algorithmic translation of expert knowledge. Our work shows that this process requires the sequencing of several transformations, including, for track definition, the choice of (1) a time-dependent signal or an already available track, (2) a statistical unit, (3) a type of track, and (4) a value or a set of values, and track aggregation is the final transformation based on (5) a track, which is performed during (6) a period of interest and involves (7) an extraction method. The formalization and documentation of these 7 items should enhance the reproducibility of studies and the sharing of features between collaborators, by removing the ambiguity about what is being calculated.

Limits

In this study, we focused on feature extraction based on expert rules and did not take into account feature extraction based on deep learning techniques [61,62]. In this case, although the aim is also to reduce the dimensionality of the source data, there is no need to interpret features, which are often abstract and designed to result in the best prediction model without being interpreted [62]. Recent advances in natural language processing [63-65] could be leveraged to automatize the extraction of relevant clinical features from clinical text [66]. Once the feature of interest has been well defined, a small annotation campaign should be conducted to fine-tune and evaluate pretrained model performances. Afterwards, the extracted feature can be integrated in our workflow as a new structured piece of information. The impressive results of large language models suggest that a few labeled examples are sufficient to fine-tune these models [67]. Three limitations must be explored before using these models. First, due to the variability of the wordings of clinical concepts, it has not been proved that a large language model can capture every targeted feature. Second, the computing intensiveness is incompatible with large-scale information retrieval. Third, the ability to conduct quick targeted annotation campaigns for precise clinical terms requires appropriate tooling and processes. We have not provided any use cases involving text. However, both tracks and features could be constructed from, for example, the presence of a symptom or the reporting of a scale in a consultation report. Such extraction from raw text raises the question of the automatic detection of specific concepts in text and the performance of the tools used for this.

Although some features, such as the length of stay, are generic and frequently used, the majority remain dependent on the study context. The period of interest and the extraction method are proxies for what is expected by the clinician or researcher, and the feature would need to be manually evaluated to ensure its validity [49].

Even if SNOMED CT (Systematized Nomenclature of Medicine - Clinical Terms) and ICD10 propose aggregate concepts, such as “Hypotension following procedure” (SNOMED CT code 16055431000119108), “Decreased mean arterial pressure” (SNOMED CT code 31013001), or “Hypotension” (ICD10 code I95), these concepts are only a part of the label of a feature, and they do not document how to compute the feature or mention the period (ie, surgery, anesthesia, intensive care unit stay, or first day of hospitalization). Standardized concepts that fully document features are yet to be defined in these terminologies.

At present, we cannot judge the generalization of our proposal. However, this study is the first to propose a standardized description of feature extraction from structured databases. The approach remains to be evaluated by comparing it with other study cases, particularly from other countries.

The next step of this project is the implementation of an R package with functions dedicated to the definition and aggregation of tracks. This package would rely on the OMOP CDM and allow reproducibility of feature extraction. Attention will need to be paid to the physical implementation of the 2 tables and, in particular, to the storage of tracks, which can be voluminous and can impact performance with regard to queries and response times. Finally, it would be relevant to implement a data mart with features arranged in columns (when they are still stored in rows in the feature table) to gain time when building tables to construct cohorts.

Conclusions

We have clarified the process of feature extraction implemented when conducting retrospective observational studies. We identified 2 transformations (track definition and track aggregation) to transform complex raw data into tracks and features. Track definition requires high expertise, but reduces the complexity of data and simplifies the reduction of time dimensionality during track aggregation.

Authors' Contributions

AL, MF, and EC contributed to study conception and design, and drafted the manuscript. All authors provided their study cases and approved the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Description of the study cases.
[DOCX File , 15 KB - medinform_v10i10e38936_app1.docx]

Multimedia Appendix 2
Standardized description of the tracks and features implemented for each study case.
[PDF File (Adobe PDF File), 100 KB - medinform_v10i10e38936_app2.pdf]

References

1. Weng C, Kahn MG. Clinical Research Informatics for Big Data and Precision Medicine. Yearb Med Inform 2016 Nov 10(1):211-218 [FREE Full text] [doi: 10.15265/IY-2016-019] [Medline: 27830253]
2. Adler-Milstein J, DesRoches CM, Kralovec P, Foster G, Worzala C, Charles D, et al. Electronic Health Record Adoption In US Hospitals: Progress Continues, But Challenges Persist. Health Aff (Millwood) 2015 Dec;34(12):2174-2180. [doi: 10.1377/hlthaff.2015.0992] [Medline: 26561387]
3. Safran C, Bloomrosen M, Hammond WE, Labkoff S, Markel-Fox S, Tang PC, Expert Panel. Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper. J Am Med Inform Assoc 2007;14(1):1-9 [FREE Full text] [doi: 10.1197/jamia.M2273] [Medline: 17077452]
4. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inform Assoc 2013 Jan 01;20(1):117-121 [FREE Full text] [doi: 10.1136/amiajnl-2012-001145] [Medline: 22955496]
5. Forrest CB, Margolis PA, Bailey LC, Marsolo K, Del Beccaro MA, Finkelstein JA, et al. PEDSnet: a National Pediatric Learning Health System. J Am Med Inform Assoc 2014;21(4):602-606 [FREE Full text] [doi: 10.1136/amiajnl-2014-002743] [Medline: 24821737]
6. McGlynn EA, Lieu TA, Durham ML, Bauck A, Laws R, Go AS, et al. Developing a data infrastructure for a learning health system: the PORTAL network. J Am Med Inform Assoc 2014;21(4):596-601 [FREE Full text] [doi: 10.1136/amiajnl-2014-002746] [Medline: 24821738]
7. Safran C. Reuse of clinical data. Yearb Med Inform 2014 Aug 15;9:52-54 [FREE Full text] [doi: 10.15265/IY-2014-0013] [Medline: 25123722]
8. De Moor G, Sundgren M, Kalra D, Schmidt A, Dugas M, Claerhout B, et al. Using electronic health records for clinical research: the case of the EHR4CR project. J Biomed Inform 2015 Feb;53:162-173 [FREE Full text] [doi: 10.1016/j.jbi.2014.10.006] [Medline: 25463966]
9. Lin K, Schneeweiss S. Considerations for the analysis of longitudinal electronic health records linked to claims data to study the effectiveness and safety of drugs. Clin Pharmacol Ther 2016 Aug 12;100(2):147-159. [doi: 10.1002/cpt.359] [Medline: 26916672]
10. Meystre SM, Lovis C, Bürkle T, Tognola G, Budrionis A, Lehmann CU. Clinical Data Reuse or Secondary Use: Current Status and Potential Future Progress. Yearb Med Inform 2017 Sep 11;26(01):38-52. [doi: 10.15265/iy-2017-007]
11. Krishnankutty B, Bellary S, Kumar NBR, Moodahadu LS. Data management in clinical research: An overview. Indian J Pharmacol 2012 Mar;44(2):168-172 [FREE Full text] [doi: 10.4103/0253-7613.93842] [Medline: 22529469]
12. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc 2013 Jan 01;20(1):144-151 [FREE Full text] [doi: 10.1136/amiajnl-2011-000681] [Medline: 22733976]
13. Safran C. Medicine based upon data. J Gen Intern Med 2013 Dec;28(12):1545-1546 [FREE Full text] [doi: 10.1007/s11606-013-2549-3] [Medline: 23838902]
14. Weiskopf NG, Bakken S, Hripcsak G, Weng C. A Data Quality Assessment Guideline for Electronic Health Record Data Reuse. EGEMS (Wash DC) 2017 Sep 04;5(1):14 [FREE Full text] [doi: 10.5334/egems.218] [Medline: 29881734]
15. Wendl C, Duftschmid G, Gezgin D, Popper N, Miksch F, Rinner C. A Web-Based Tool to Evaluate Data Quality of Reused Health Data Assets. Stud Health Technol Inform 2017;236:204-210. [Medline: 28508797]
16. Wang Z, Dagtas S, Talburt J, Baghal A, Zozus M. Rule-Based Data Quality Assessment and Monitoring System in Healthcare Facilities. Stud Health Technol Inform 2019;257:460-467. [Medline: 30741240]

17. Rahm E, Do HH. Data Cleaning: Problems and Current Approaches. Better Evaluation. URL: https://www.betterevaluation.org/sites/default/files/data_cleaning.pdf [accessed 2022-09-24]
18. Weng C. Clinical data quality: a data life cycle perspective. Biostat Epidemiol 2020;4(1):6-14. [doi: 10.1080/24709360.2019.1572344] [Medline: 32258941]
19. Murff HJ, FitzHenry F, Matheny ME, Gentry N, Kotter KL, Crimin K, et al. Automated identification of postoperative complications within an electronic medical record using natural language processing. JAMA 2011 Aug 24;306(8):848-855. [doi: 10.1001/jama.2011.1204] [Medline: 21862746]
20. Smischney NJ, Velagapudi VM, Onigkeit JA, Pickering BW, Herasevich V, Kashyap R. Derivation and validation of a search algorithm to retrospectively identify mechanical ventilation initiation in the intensive care unit. BMC Med Inform Decis Mak 2014 Jun 25;14:55 [FREE Full text] [doi: 10.1186/1472-6947-14-55] [Medline: 24965680]
21. Tien M, Kashyap R, Wilson GA, Hernandez-Torres V, Jacob AK, Schroeder DR, et al. Retrospective Derivation and Validation of an Automated Electronic Search Algorithm to Identify Post Operative Cardiovascular and Thromboembolic Complications. Appl Clin Inform 2015;6(3):565-576 [FREE Full text] [doi: 10.4338/ACI-2015-03-RA-0026] [Medline: 26448798]
22. Lamer A, Jeanne M, Marcilly R, Kipnis E, Schiro J, Logier R, et al. Methodology to automatically detect abnormal values of vital parameters in anesthesia time-series: Proposal for an adaptable algorithm. Comput Methods Programs Biomed 2016 Jun;129:160-171. [doi: 10.1016/j.cmpb.2016.01.004] [Medline: 26817405]
23. Gabel E, Hofer IS, Satou N, Grogan T, Shemin R, Mahajan A, et al. Creation and Validation of an Automated Algorithm to Determine Postoperative Ventilator Requirements After Cardiac Surgery. Anesth Analg 2017 May;124(5):1423-1430. [doi: 10.1213/ANE.0000000000001997] [Medline: 28431419]
24. Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. AI Magazine 1996;17(3):37-54 [FREE Full text] [doi: 10.1609/aimag.v17i3.1230]
25. Lamer A, Jeanne M, Ficheur G, Marcilly R. Automated Data Aggregation for Time-Series Analysis: Study Case on Anaesthesia Data Warehouse. Stud Health Technol Inform 2016;221:102-106. [Medline: 27071886]
26. Price A, Caciula A, Guo C, Lee B, Morrison J, Rasmussen A, et al. DEvis: an R package for aggregation and visualization of differential expression data. BMC Bioinformatics 2019 Mar 04;20(1):110 [FREE Full text] [doi: 10.1186/s12859-019-2702-z] [Medline: 30832568]
27. Chazard E, Ficheur G, Caron A, Lamer A, Labreuche J, Cuggia M, et al. Secondary Use of Healthcare Structured Data: The Challenge of Domain-Knowledge Based Extraction of Features. Stud Health Technol Inform 2018;255:15-19. [Medline: 30306898]
28. Pasma W, Peelen LM, van Buuren S, van Klei WA, de Graaff JC. Artifact Processing Methods Influence on Intraoperative Hypotension Quantification and Outcome Effect Estimates. Anesthesiology 2020 Apr;132(4):723-737 [FREE Full text] [doi: 10.1097/ALN.0000000000003131] [Medline: 32022770]
29. Breil B, Kenneweg J, Fritz F, Bruland P, Doods D, Trinczek B, et al. Multilingual Medical Data Models in ODM Format: A Novel Form-based Approach to Semantic Interoperability between Routine Healthcare and Clinical Research. Appl Clin Inform 2012;3(3):276-289 [FREE Full text] [doi: 10.4338/ACI-2012-03-RA-0011] [Medline: 23620720]
30. Pivovarov R, Albers DJ, Sepulveda JL, Elhadad N. Identifying and mitigating biases in EHR laboratory tests. J Biomed Inform 2014 Oct;51:24-34 [FREE Full text] [doi: 10.1016/j.jbi.2014.03.016] [Medline: 24727481]
31. Krumm R, Semjonow A, Tio J, Duhme H, Bürkle T, Haier J, et al. The need for harmonized structured documentation and chances of secondary use - results of a systematic analysis with automated form comparison for prostate and breast cancer. J Biomed Inform 2014 Oct;51:86-99 [FREE Full text] [doi: 10.1016/j.jbi.2014.04.008] [Medline: 24747879]
32. Dhombres F, Bodenreider O. Interoperability between phenotypes in research and healthcare terminologies--Investigating partial mappings between HPO and SNOMED CT. J Biomed Semantics 2016;7:3 [FREE Full text] [doi: 10.1186/s13326-016-0047-3] [Medline: 26865946]
33. Dugas M, Neuhaus P, Meidt A, Doods J, Storck M, Bruland P, et al. Portal of medical data models: information infrastructure for medical research and healthcare. Database (Oxford) 2016;2016:bav121 [FREE Full text] [doi: 10.1093/database/bav121] [Medline: 26868052]
34. Xu Y, Zhou X, Suehs BT, Hartzema AG, Kahn MG, Moride Y, et al. A Comparative Assessment of Observational Medical Outcomes Partnership and Mini-Sentinel Common Data Models and Analytics: Implications for Active Drug Safety Surveillance. Drug Saf 2015 Aug;38(8):749-765. [doi: 10.1007/s40264-015-0297-5] [Medline: 26055920]
35. Garza M, Del Fiol G, Tenenbaum J, Walden A, Zozus MN. Evaluating common data models for use with a longitudinal community registry. J Biomed Inform 2016 Dec;64:333-341 [FREE Full text] [doi: 10.1016/j.jbi.2016.10.016] [Medline: 27989817]
36. Klann JG, Abend A, Raghavan VA, Mandl KD, Murphy SN. Data interchange using i2b2. J Am Med Inform Assoc 2016 Sep;23(5):909-915 [FREE Full text] [doi: 10.1093/jamia/ocv188] [Medline: 26911824]
37. Hume S, Aerts J, Sarnikar S, Huser V. Current applications and future directions for the CDISC Operational Data Model standard: A methodological review. J Biomed Inform 2016 Apr;60:352-362 [FREE Full text] [doi: 10.1016/j.jbi.2016.02.016] [Medline: 26944737]

38. Liyanage H, Liaw S, Jonnagaddala J, Hinton W, de Lusignan S. Common Data Models (CDMs) to Enhance International Big Data Analytics: A Diabetes Use Case to Compare Three CDMs. Stud Health Technol Inform 2018;255:60-64. [Medline: 30306907]
39. Stang PE, Ryan PB, Racoosin JA, Overhage JM, Hartzema AG, Reich C, et al. Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. Ann Intern Med 2010 Nov 02;153(9):600-606. [doi: 10.7326/0003-4819-153-9-201011020-00010] [Medline: 21041580]
40. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc 2012;19(1):54-60 [FREE Full text] [doi: 10.1136/amiajnl-2011-000376] [Medline: 22037893]
41. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform 2015;216:574-578 [FREE Full text] [Medline: 26262116]
42. Observational Health Data Sciences and Informatics. URL: https://www.ohdsi.org/ [accessed 2019-05-03]
43. Areas of Focus. Observational Health Data Sciences and Informatics. URL: https://www.ohdsi.org/who-we-are/areas-of-focus/ [accessed 2022-07-07]
44. Suchard MA, Schuemie MJ, Krumholz HM, You SC, Chen R, Pratt N, et al. Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: a systematic, multinational, large-scale analysis. Lancet 2019 Nov 16;394(10211):1816-1826 [FREE Full text] [doi: 10.1016/S0140-6736(19)32317-7] [Medline: 31668726]
45. Burn E, You SC, Sena AG, Kostka K, Abedtash H, Abrahão MTF, et al. Deep phenotyping of 34,128 adult patients hospitalised with COVID-19 in an international network study. Nat Commun 2020 Oct 06;11(1):5009 [FREE Full text] [doi: 10.1038/s41467-020-18849-z] [Medline: 33024121]
46. Belenkaya R, Gurley MJ, Golozar A, Dymshyts D, Miller RT, Williams AE, et al. Extending the OMOP Common Data Model and Standardized Vocabularies to Support Observational Cancer Research. JCO Clin Cancer Inform 2021 Jan;5:12-20 [FREE Full text] [doi: 10.1200/CCI.20.00079] [Medline: 33411620]
47. Conversion of Diagnosis and Chemotherapy Data in Electronic Health Records to Episode-based Oncology Extension of OMOP-CDM. Observational Health Data Sciences and Informatics. URL: https://www.ohdsi.org/2019-us-symposium-showcase-12/ [accessed 2021-10-22]
48. Warner JL, Dymshyts D, Reich CG, Gurley MJ, Hochheiser H, Moldwin ZH, et al. HemOnc: A new standard vocabulary for chemotherapy regimen representation in the OMOP common data model. J Biomed Inform 2019 Aug;96:103239 [FREE Full text] [doi: 10.1016/j.jbi.2019.103239] [Medline: 31238109]
49. Jeon H, You SC, Kang SY, Seo SI, Warner JL, Belenkaya R, et al. Characterizing the Anticancer Treatment Trajectory and Pattern in Patients Receiving Chemotherapy for Cancer Using Harmonized Observational Databases: Retrospective Study. JMIR Med Inform 2021 Apr 06;9(4):e25035 [FREE Full text] [doi: 10.2196/25035] [Medline: 33720842]
50. Boudemaghe T, Belhadj I. Data Resource Profile: The French National Uniform Hospital Discharge Data Set Database (PMSI). Int J Epidemiol 2017 Apr 01;46(2):392-392d. [doi: 10.1093/ije/dyw359] [Medline: 28168290]
51. Scailteux L, Droitcourt C, Balusson F, Nowak E, Kerbrat S, Dupuy A, et al. French administrative health care database (SNDS): The value of its enrichment. Therapie 2019 Apr;74(2):215-223. [doi: 10.1016/j.therap.2018.09.072] [Medline: 30392702]
52. Lamer A, Jeanne M, Vallet B, Ditilyeu G, Delaby F, Tavernier B, et al. Development of an anesthesia data warehouse: Preliminary results. IRBM 2013 Dec;34(6):376-378. [doi: 10.1016/j.irbm.2013.09.005]
53. Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 2000 Jun 13;101(23):E215-E220. [doi: 10.1161/01.cir.101.23.e215] [Medline: 10851218]
54. Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016 May 24;3:160035 [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
55. French A, Taylor LK, Lemke MR. Chapter 6 - Task analysis. In: Privitera MB, editor. Applied Human Factors in Medical Device Design. Cambridge, MA: Academic Press; 2019:63-81.
56. Popoff B, Besnier E, Dureuil B, Veber B, Clavier T. Effect of early hyperoxemia on mortality in mechanically ventilated septic shock patients according to Sepsis-3 criteria: analysis of the MIMIC-III database. Eur J Emerg Med 2021 Dec 01;28(6):469-475. [doi: 10.1097/MEJ.0000000000000854] [Medline: 34285171]
57. Laurent G, Moussa MD, Cirenei C, Tavernier B, Marcilly R, Lamer A. Development, implementation and preliminary evaluation of clinical dashboards in a department of anesthesia. J Clin Monit Comput 2021 May;35(3):617-626 [FREE Full text] [doi: 10.1007/s10877-020-00522-x] [Medline: 32418147]
58. Laroche M, Charmes J, Merle L. Potentially inappropriate medications in the elderly: a French consensus panel list. Eur J Clin Pharmacol 2007 Aug;63(8):725-731. [doi: 10.1007/s00228-007-0324-2] [Medline: 17554532]
59. Chazard E, Boudry A, Beeler PE, Dalleur O, Hubert H, Tréhou E, et al. Towards The Automated, Empirical Filtering of Drug-Drug Interaction Alerts in Clinical Decision Support Systems: Historical Cohort Study of Vitamin K Antagonists. JMIR Med Inform 2021 Jan 20;9(1):e20862 [FREE Full text] [doi: 10.2196/20862] [Medline: 33470938]

60. EPISODE. Observational Health Data Sciences and Informatics. URL: https://ohdsi.github.io/CommonDataModel/cdm54.html#EPISODE [accessed 2021-10-21]
61. Liang H, Sun X, Sun Y, Gao Y. Text feature extraction based on deep learning: a review. EURASIP J Wirel Commun Netw 2017;2017(1):211 [FREE Full text] [doi: 10.1186/s13638-017-0993-1] [Medline: 29263717]
62. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015 May 28;521(7553):436-444. [doi: 10.1038/nature14539] [Medline: 26017442]
63. Peters M, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, et al. Deep contextualized word representations. arXiv. 2018. URL: https://arxiv.org/abs/1802.05365 [accessed 2022-09-24]
64. Devlin J, Chang M, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv. 2019. URL: https://arxiv.org/abs/1810.04805 [accessed 2022-09-24]
65. Brown T, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language Models are Few-Shot Learners. arXiv. 2020. URL: https://arxiv.org/abs/2005.14165 [accessed 2022-09-24]
66. Alsentzer E, Murphy J, Boag W, Weng W, Jin D, Naumann T, et al. Publicly Available Clinical BERT Embeddings. arXiv. 2019. URL: https://arxiv.org/abs/1904.03323 [accessed 2022-09-24]
67. Copara J, Knafou J, Naderi N, Moro C, Ruch P, Teodoro D. Contextualized French Language Models for Biomedical Named Entity Recognition. HAL Archives. URL: https://hal.archives-ouvertes.fr/hal-02784740 [accessed 2022-07-18]

Abbreviations

ATC: Anatomical Therapeutic Chemical
CDM: common data model
HAT: hierarchical analysis of the task
ICD10: International Statistical Classification of Diseases and Related Health Problems, 10th Revision
OHDSI: Observational Health Data Sciences and Informatics
OMOP: Observational Medical Outcomes Partnership
SNDS: Système National des Données de Santé (French national claims database)
SNOMED CT: Systematized Nomenclature of Medicine - Clinical Terms

Edited by C Lovis, J Hefner; submitted 22.04.22; peer-reviewed by M Sedlmayr, FM Calisto, E Sylvestre; comments to author 13.06.22; revised version received 19.07.22; accepted 11.08.22; published 17.10.22.

Please cite as:
Lamer A, Fruchart M, Paris N, Popoff B, Payen A, Balcaen T, Gacquer W, Bouzillé G, Cuggia M, Doutreligne M, Chazard E Standardized Description of the Feature Extraction Process to Transform Raw Data Into Meaningful Information for Enhancing Data Reuse: Consensus Study
JMIR Med Inform 2022;10(10):e38936
URL: https://medinform.jmir.org/2022/10/e38936
doi:10.2196/38936
PMID:36251369

Original Paper

A Recurrent Neural Network Model for Predicting Activated Partial Thromboplastin Time After Treatment With Heparin: Retrospective Study

Sebastian Daniel Boie[1], PhD; Lilian Jo Engelhardt[1,2], MD; Nicolas Coenen[2], MD; Niklas Giesa[1], MSc; Kerstin Rubarth[1,3], MSc; Mario Menk[1,2], MD; Felix Balzer[1], PhD, MD

[1]Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Institute of Medical Informatics, Berlin, Germany

[2]Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Department of Anesthesiology and Intensive Care Medicine, Berlin, Germany

[3]Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Institute of Biometry and Clinical Epidemiology, Berlin, Germany

Corresponding Author:
Sebastian Daniel Boie, PhD
Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin
Institute of Medical Informatics
Charitéplatz 1
Berlin, 10117
Germany
Phone: 49 30 450580877
Email: Sebastian-Daniel.Boie@charite.de

Abstract

Background: Anticoagulation therapy with heparin is a frequent treatment in intensive care units and is monitored by activated partial thromboplastin clotting time (aPTT). It has been demonstrated that reaching an established anticoagulation target within 24 hours is associated with favorable outcomes. However, patients respond to heparin differently and reaching the anticoagulation target can be challenging. Machine learning algorithms may potentially support clinicians with improved dosing recommendations.

Objective: This study evaluates a range of machine learning algorithms on their capability of predicting the patients' response to heparin treatment. In this analysis, we apply, for the first time, a model that considers time series.

Methods: We extracted patient demographics, laboratory values, dialysis and extracorporeal membrane oxygenation treatments, and scores from the hospital information system. We predicted the numerical values of aPTT laboratory values 24 hours after continuous heparin infusion and evaluated 7 different machine learning models. The best-performing model was compared to recently published models on a classification task. We considered all data before and within the first 12 hours of continuous heparin infusion as features and predicted the aPTT value after 24 hours.

Results: The distribution of aPTT in our cohort of 5926 hospital admissions was highly skewed. Most patients showed aPTT values below 75 s, while some outliers showed much higher aPTT values. A recurrent neural network that consumes a time series of features showed the highest performance on the test set.

Conclusions: A recurrent neural network that uses time series of features instead of only static and aggregated features showed the highest performance in predicting aPTT after heparin treatment.

(JMIR Med Inform 2022;10(10):e39187) doi:10.2196/39187

KEYWORDS

machine learning; health care; recurrent neural network; heparin; activated partial thromboplastin time (aPTT); deep learning; ICU; critical care

Introduction

Thromboembolic complications are associated with increased mortality [1,2]. Risk factors for deep venous thrombosis and pulmonary embolism include, for example, immobility, malignancy, higher age, and a history of thromboembolism [3,4]. Anticoagulation by drugs is applied either prophylactically to prevent thromboembolism [5] or therapeutically to treat existing thromboembolic complications, which reduces mortality [6].

In perioperative normal care wards, prophylactic and therapeutic anticoagulation is frequently performed subcutaneously by low–molecular weight heparins [5]. In the perioperative setting, prophylactic anticoagulation is indicated in patients with intermediate or high risk for thromboembolism. This includes, for example, most trauma surgeries, elective orthopedic surgeries with consecutive immobility of the lower limbs, and major abdominal or thoracic surgery, particularly in the presence of malignant and inflammatory processes [5].

In critical illness, the risk for venous thromboembolism is increased in almost all patients due to the combination of general risk factors related to chronic disease and intensive care unit (ICU)–associated risk factors, including sedation, immobility, or central venous catheters [7]. In intensive care, prophylactic or therapeutic anticoagulation is regularly applied intravenously by continuous unfractionated heparin, particularly during renal failure or hemodynamic instability [8]. The short half-life of the anticoagulant and the possibility of antagonizing heparin with protamine are advantages of unfractionated heparin in these vulnerable patients [9]. However, poor controllability is an issue. Consequently, overdosing with hemorrhagic or underdosing with thrombotic complications may occur [10]. Hence, therapeutic unfractionated heparin application requires monitoring. The dosing of unfractionated heparin is performed by determination of activated partial thromboplastin time (aPTT) in patients' blood [11]. Based on older studies, the pursued aPTT target is approximately a 1.5 to 2.5-times prolongation of the reference clotting time [11-13] although individual targets are usually defined. Achieving the aPTT target within 24 hours has been associated with increased survival in patients with pulmonary embolism [6]. However, due to patient- and disease-related variations, achieving the aPTT target within 24 hours is challenging.

Nowadays, big data sets are generated by digital patient data management systems in ICU routine. Machine learning (ML) approaches that include individual information from large data sets may help to predict aPTT at an earlier stage than can routine blood sampling. Previous results of applying ML to predict aPTT show great promise [14-17]. Some authors [16,17] consider the numerical value of aPTT and consequently the prediction of aPTT as a regression task. We prefer the prediction of the numerical value since it makes no assumption of the aPTT target range. However, most recent literature on similar-sized data sets consider aPTT after heparin treatment as a multiclass prediction with 3 distinct ranges (subtherapeutic, therapeutic, or supratherapeutic) [14,15,18].

In previous model comparison studies [15,16,18], it has been demonstrated that artificial neural networks show the highest performance on aPTT prediction tasks.

Recently, a systematic review of ML approaches on predicting aPTT after heparin administration highlighted that still multiple innovations are required before ML-assisted heparin dosing is ready for clinical practice [19].

We compared multiple ML models on our patient cohort and are, to our knowledge, the first to apply a recurrent neural network model that takes the dynamics of variables in the form of time series into account. At the outset of the study, we specified inclusion criteria that resulted in 5926 distinct hospital admissions. On this cohort, we trained and evaluated multiple ML models on the aPTT prediction task. To allow comparison of the recurrent neural network model with previously published models [14,15,18], we subsequently used our model in a classification setup.

The aim of this analysis is to evaluate whether ML models can accurately predict subsequent aPTT measurements well (12 hours) in advance. In the future, data-driven approaches could help clinicians to adjust heparin dosing to improve time in the target range aPTT after 24 hours.

Methods

Data Selection Criteria

The database system for surgical and intensive care patients at Charité – Universitätsmedizin Berlin (Charité) was first adopted in 2012 and over time rolled out to all ICUs. Since we extracted data in November 2021, we considered a time period from 2012 to October 31, 2021. We selected patients and hospital admissions that satisfied the following inclusion criteria: at least 18 years old at the beginning of treatment, received a minimum of 1000 IU of heparin, received some of the heparin as continuous infusion, had at least a single aPTT measurement after 12 hours and before 36 hours after the intravenous treatment commenced, and had weight and height documented (within reasonable limits: height between 25 cm to 250 cm, weight between 3 kg to 300 kg).

Ethics Approval

Ethics approval for this study was obtained by the Charité ethics committee (vote #EA4/241/21).

Feature Selection and Prediction Targets

We extracted patient characteristics (age, gender, height, weight), laboratory values (aPTT, bilirubin, C-reactive protein, creatinine, quick value, platelet count), whether patients received dialysis or a form of extracorporeal membrane oxygenation (ECMO) treatment, and routinely collected scores (therapeutic intervention scoring system 10 [TISS-10], simplified acute physiology score [SAPS-II], sequential organ failure assessment [SOFA], acute physiology and chronic health evaluation II [APACHE II]) from the hospital information system. Furthermore, we extracted the time of the start and end of each heparin dosing, concentration, and administration rate. Heparin can be administered as a bolus or as a continuous infusion. All

data were restricted to the time period 7 days prior to treatment to 36 hours after treatment started.

Our goal was to predict the aPTT 24 hours after initiation of continuous heparin treatment. However, not all patients had a laboratory measurement exactly 24 hours after the treatment with heparin. Thus, any aPTT measurement between 12 to 36 hours after heparin treatment began was accepted as the prediction target. In case multiple values were recorded between 12 hours and 36 hours, we chose the value that was closest to 24 hours after continuous treatment started. Consequently, only values that were taken before or within 12 hours after continuous heparin treatment commenced were available as features for the model (including any aPTT measurement in that time frame). Hospital stays were left aligned, and the start of the continuous intravenous heparin delivery corresponded to time zero.

Handling of Missing Data

The data we used for our study were collected during routine care and were not of uniform quality across all hospital admissions. A typical problem when using retrospective data for ML is missing observations [20-22]. This problem is exacerbated for the recurrent neural network, as it expects an input for every feature every 2 hours.

The static values of gender, age, height, and weight had no missing values and were replicated for every timestamp. The one-hot–encoded variables, including ECMO treatment, dialysis, bolus delivery of heparin, and continuous delivery of heparin, were set to 0 if no other value was recorded for a given timestamp. Other features (eg, laboratory measurements and scores) were filled in a 2-step process as follows: (1) If a previous value was recorded within 7 days prior to continuous heparin treatment, those values were forward filled; (2) Any still missing values were replaced by the mean across the training population.

Only using the above 2-step procedure discards information about which measurement is from the patient at the given timestamp. Since it has been shown that the missing pattern can be informative [23], we included an "indicator" variable for each variable filled in the 2-step process that is 1 if the value was measured at the given timestamp and 0 if it was imputed.

Together with the indicator variables, each model sees 35 different input variables.

The recurrent neural network, thus, may see time series between $t = –168$ (7 days prior to continuous heparin delivery) to $t= 12$. In general, however, patients' time series are not of the same length.

Models and Variable Encoding

The input data consisted of numerical and categorical variables. Categorical variables (gender, ECMO treatment, dialysis treatment, continuous heparin administration, bolus heparin administration) were one-hot encoded. Each option for a categorical variable resulted in 1 input dimension that could either be 1 or 0. One-hot–encoded variables were not further scaled and were directly used as input features.

Other numerical variables were standardized before being fed into the model. Mean and SD were estimated only on the training data set.

We compared 6 models that take a single value per feature and 1 model that takes the entire time series of features. Some features were constant over the course of treatment (age, gender, height, and weight), while the other features changed frequently. Models that take a single value per feature received the last-observed value before the 12-hour cutoff. The recurrent neural network received time series, resampled to 2-hour intervals, for each feature. If multiple measurements were taken within 2 hours, those values were replaced by the mean over this 2-hour window. Static variables were repeated for each timestamp. The prediction target (a single aPTT measurement) is log-transformed during model training. The log transformation is discussed in the Results section. All model parameters are optimized on the mean-squared error (MSE) loss function. Additionally, we evaluated the mean absolute error and the explained variance for each model.

The 6 regression models were linear regression, elastic net, generalized linear model, support vector machine regression (SVR), K-nearest neighbor regression (KNN), and regression trees. We optimized hyperparameters using a grid search with 5-fold cross-validation. For the cross-validation, training and validation data were combined. The hyperparameter grids are shown in Table 1.

The models, cross-validation, and the grid search routine were from the scikit-learn package [24] and implemented in Python (The Python Software Foundation).

Table 1. Hyperparameters for each static model.

Model	Hyperparameters
Linear regression	None
Elastic net	$\alpha = (10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 2, 3)$ $_{L1}$ratio = (0, 0.1, … 1.0)
GLM[a]	Power = (0, 1, 2, 3) $\alpha = (10^{-2}, 10^{-1}, 1, 2, 3)$
SVR[b]	Kernel = (“linear,” “poly,” “rbf,” “sigmoid”) Degree = (2, 3, 4, 5, 6)
KNN[c]	K = (2, 3, 4, 5, 6, 7, 8, 9, 10) Weights = (“uniform,” “distance”)
Regression trees	Max_depth = (2, 3, 4, 5, unlimited) Min_samples_split = (2, 3, 4, 5, 6) Min_samples_leaf = (1, 2, 3, 4, 5)

[a]GLM: generalized linear model.

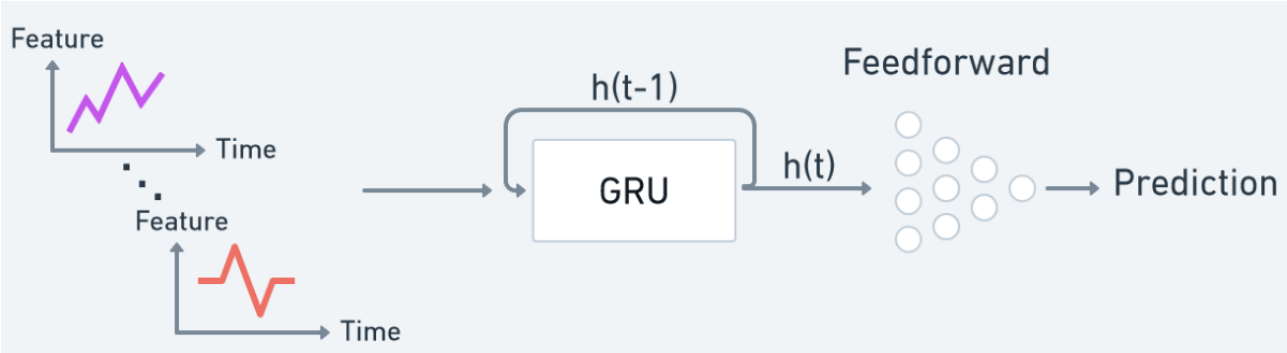
[b]SVR: support vector machine regression.

[c]KNN: K-nearest neighbor regression.

Recurrent Neural Network Model

This model consists of a gated recurrent unit (GRU), which can process a time series of arbitrary length and a fully connected network that uses the output of the GRU as input. Since we are only interested in predicting a single value, only the last output of the GRU is fed into a 3-layer fully connected model. No activation function is used between the output of the GRU and the first fully connected layer. The outputs of the 2 fully connected layers have rectified linear unit activation functions [25], and the final layer has no activation function. A schematic overview can be seen in Figure 1.

Figure 1. Schematic overview of input features, recurrent neural network, and feedforward network. GRU: gated recurrent unit.



As for the previously described models, the recurrent neural network was optimized on the MSE. For experiments with the recurrent neural network, weights were optimized on the training set, and the results between experiments were compared on the validation set. We used the Adam optimizer with L2 penalty [26]. For each experiment, we chose weights with the lowest error on the validation set, which may occur before the maximum number of epochs are reached.

This model is significantly more costly to train compared to “static” models. Therefore, we did not perform a systematic hyperparameter optimization but ran several experiments with different hyperparameters and handpicked the best set of hyperparameters, which are shown in the Results section. Hyperparameters for the GRU submodel are hidden size (n=1, 2, 3, …), bidirectional connection (True, False), and the number of layers (n=1, 2, 3, …).

The 3-layered fully connected submodel had the number of neurons in each layer as 3 hyperparameters. Hyperparameters related to the training are the learning rate, L2 penalty, and the maximum number of epochs.

Patients have a different number of inputs per feature, since they receive their continuous heparin treatment at different times within their hospital stay. Thus, for training, we are limited to a batch size of 1 but accumulate multiple batches before weights are updated. To combat overfitting, we used an L2 penalty on the weights in the fully connected part of the model and chose weights on the epoch with the highest performance on the validation set.

All models and training scripts are available on github [27].

Classification Models

To phrase aPTT prediction as a classification task, we used the 3 ranges first introduced by Ghassemi et al [14] of

subtherapeutic for values below 60 s, therapeutic for values between 60 s to 100 s, and supratherapeutic for values above 100 s for the aPTT measurements. We compared our GRU model to the logistic regression model from Ghassemi et al [14] and the feedforward neural networks models by Su et al [15] and Li et al [18]. All parameters were taken from the reference literature for the respective model. For the feedforward networks from Su et al [15] and Li et al [18], we used cross-entropy [28] as a loss function with early stopping since the loss functions are not mentioned in the references.

The 3 classification models are retrained on the training split and receive the last value of each feature before the 12-hour cutoff in the same manner as the "static" regression models. The GRU is not retrained on the classification task, but the numeric predictions are binned into the 3 ranges post hoc. We evaluated the models on macroaveraged precision, macroaveraged recall, macroaveraged F_1-score, and accuracy [29].

Results

Patient Cohort

A flow diagram of consecutively applied filter criteria (specified in the methods section) to the entire patient cohort is shown in Figure 2. The selection criteria resulted in 5926 hospital admissions from a total of 5742 unique patients. Given that fewer than 4% of admissions occurred for previously admitted patients, we considered hospital admissions to be independent events. Basic patient characteristics and missing values are documented in Table 2.

Before model training or parameter estimation for mean and SD were performed, the admissions were split into training (n=3800), validation (n=945), and test (n=1181) samples. We ensured that different admissions by the same patient were in the same fold.

Figure 2. Flow diagram of unique patients and admissions that satisfy the specified inclusion criteria. aPTT: activated partial thromboplastin time; IV: intravenous line.

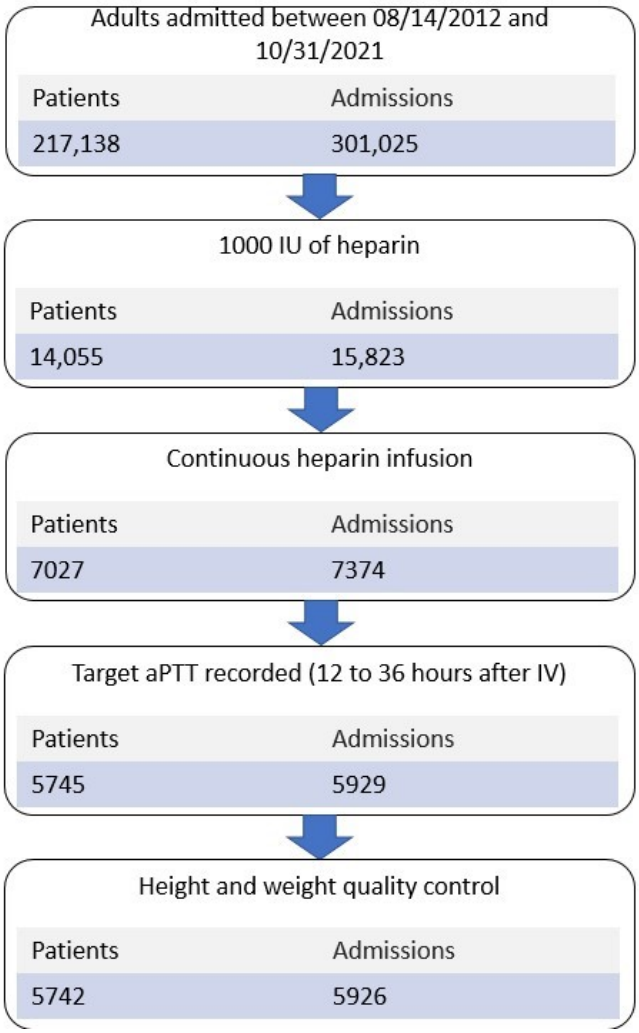


Table 2. Basic characteristics of the study cohort. The third column indicates how many patients do not have a single measurement during the hospital admission.

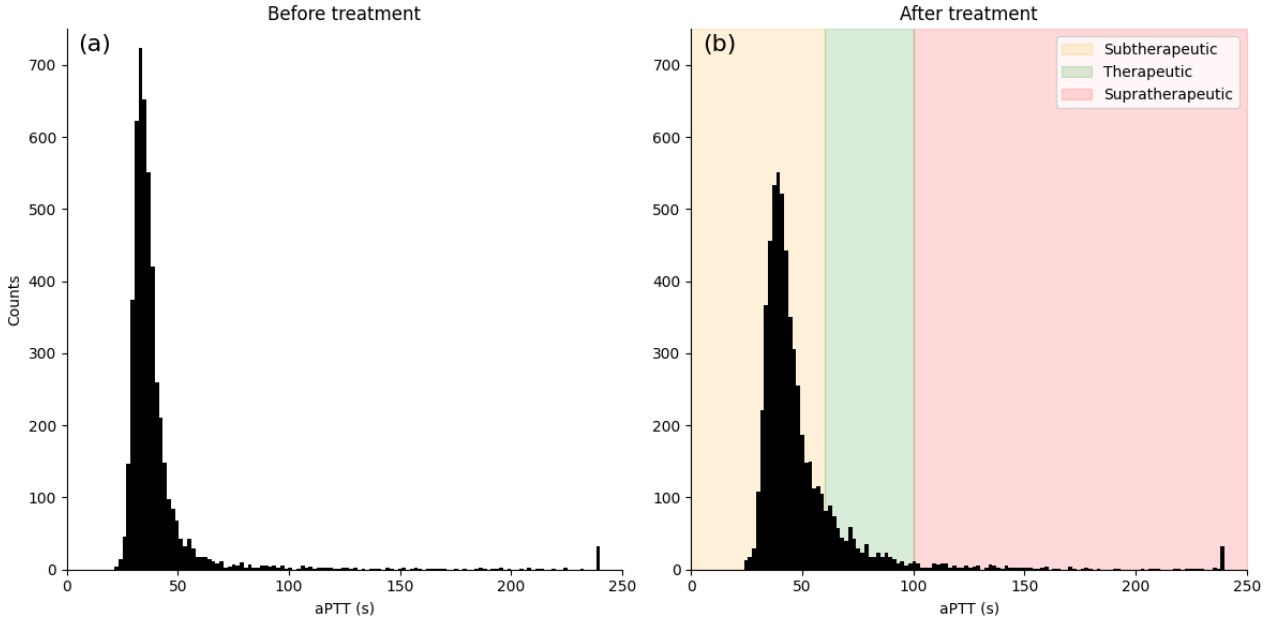
Feature	Value (N=5926)	Patients missing for entire stay, n (%)
Age (years), median (IQR)	70.62 (60.95-77.74)	0 (0)
Gender, n (%)		0 (0)
Female	1910 (32)	N/A[a]
Male	4016 (68)	N/A
Height (cm), median (IQR)	172 (164-178)	0 (0)
Weight (kg), median (IQR)	77 (66-90)	0 (0)
SOFA[b], median (IQR)	5 (2-8)	442 (7.46)
SAPS II[c], median (IQR)	36 (27-47)	449 (7.58)
APACHE II[d], median (IQR)	17 (12-23)	525 (8.86)
TISS-10[e], median (IQR)	10 (5-15)	5755 (97.11)
Dialysis, n (%)	449 (7.57)	0 (0)
ECMO[f], n (%)	76 (1.28)	0 (0)
aPTT[g] (s), median (IQR)	42.6 (36.1-54.6)	0 (0)
Bilirubin (mg/dL), median (IQR)	0.6 (0.35-1.24)	2529 (42.69)
CRP[h] (mg/L), median (IQR)	56.2 (18.6-118.8)	1782 (30.07)
Gfr[i] (count), median (IQR)	67 (39-90)	71 (1.20)
Creatinine (mg/dL), median (IQR)	1.01 (0.74-1.56)	32 (0.54)
Quick value (%), median (IQR)	76 (64-87)	17 (0.29)
Platelet count (per nL), median (IQR)	204 (139-292)	19 (0.32)
Total heparin administered (IU), median (IQR)	32398 (9500-90000)	0 (0)

[a]N/A: not applicable.
[b]SOFA: sequential organ failure assessment.
[c]SAPS II: simplified acute physiology score II.
[d]APACHE II: acute physiology and chronic health evaluation II.
[e]TISS-10: therapeutic intervention scoring system 10.
[f]ECMO: extracorporeal membrane oxygenation.
[g]aPTT: activated partial thromboplastin time.
[h]CRP: C-reactive protein.
[i]Gfr: glomerular filtration rate.

Distribution of aPTT Values

A histogram of measured aPTT before and after treatment is shown in Figure 3. In our cohort, both aPTT distributions before and after heparin treatment are narrowly peaked with a heavy tail. Values above 100 s occur very rarely. Small peaks are visible at 240 s where the laboratory reports some values as >240 s, which is mapped to 240 s.

Figure 3. Histogram of aPTT values before treatment (a) and after treatment (b) with intravenous heparin. The histogram was obtained through binning, using 120 bins between minimal and maximal values. Shaded regions indicate regions identified in Ghassemi et al [14] and Su et al [15]. aPTT: activated partial thromboplastin time.



The effect of heparin treatment on the entire cohort is clearly seen by the shift of the distribution. The difference in means is 8.64 s (95% CI 7.72-9.56; P<.001). The first 4 moments of the distribution of aPTT at t=0 and at t=24 are documented in Table 3. The mean aPTT value is higher after continuous heparin delivery compared to before treatment. Skew and kurtosis (while smaller after treatment) quantifiably indicate that the aPTT distribution is not symmetric and has a heavy tail. This fact makes the prediction of aPTT challenging. To make the learning task easier for our models, we log-transform the target variable to reduce skew and kurtosis. In effect, this makes "rare" events in the original distribution easier to predict.

The distribution that we observed in the Charité cohort contrasts with the aPTT values that are documented by other authors. Su et al [15] and Ghassemi et al [14] base their modeling studies on the Medical Information Mart for Intensive Care (MIMIC) II/III and eICU databases. The distribution of aPTT on the eICU database [15] is more heavy tailed than is the MIMIC cohort, however, less so than is our cohort. The 3 treatment categories reported in those works are indicated as shaded regions in Figure 3b. However, we do not classify our cohort into these categories but treat the prediction of aPTT after treatment as a regression problem.

Table 3. Statistical description of the binned distribution of aPTT values before continuous heparin treatment (t=0), 24 hours after continuous treatment commenced (t=24), and the log-transformed distribution after 24 hours.

	aPTT[a] (t=0)	aPTT (t=24)	Log (aPTT [t=24])
Observations, n	4850	5926	5926
Mean	40.64	49.28	3.83
Variance	561.55	608.19	0.11
Skew	6.11	4.74	1.91
Kurtosis	42.93	26.71	5.37

[a]aPTT: activated partial thromboplastin time.

Model Comparisons

In this section, the results of comparing 7 different models on the prediction of aPTT (see Table 4) are shown. Models 1-6 received only the last-measured values of each input feature before the 12-hour cutoff. We optimized hyperparameters for each model using a grid search and 5-fold cross-validation. The reported results are based on the test data that was not included in the 5 folds. A full description of the used grids appears in the Methods section. The best parameters for Models 1-6 are documented in Multimedia Appendix 1.

Model 7 (recurrent neural network) consumes the entire time series, resampled to 2-hour timestamps, for each input feature. We experimented also with resampling to 1-hour time steps and 4-hour time steps and found that the performance was similar (see Multimedia Appendix 1 for numerical results).

It is the most complex model in the comparison and ingests data from up to 7 days before continuous treatment to 12 hours after continuous treatment is administered. A systematic hyperparameter optimization for Model 7 was not performed;

hence, we are underestimating the performance of the recurrent neural network in comparison to other models.

However, the recurrent neural network model achieved the highest score on the explained variance and MSE metrics. It ranked second to the SVR model on the mean absolute error (which penalizes outliers less than does the MSE). The SVR models ranked second to the recurrent neural network model on explained variance and MSE.

CIs were obtained by taking 1000 random samples of the same size as the test set, with replacement. Given that the distribution had a small number of large outliers, which had a significant effect on the quantity of interest, the CIs are wide.

Table 4. Comparison of different models for explained variance (higher is better), mean-squared error (lower is better), and mean absolute error (lower is better) obtained by resampling 1000 samples from the test set.

	Model	Explained variance	MSE[a]	MAE[b]
1	Linear regression, test set value, (95% CI)	0.163 (0.115-0.211)	0.487 (0.425-0.556)	0.474 (0.45-0.497)
2	Elastic net regression	0.168 (0.124-0.214)	0.484 (0.433-0.554)	0.474 (0.453-0.497)
3	GLM[c]	0.169 (0.121-0.21)	0.484 (0.422-0.556)	0.473 (0.450-0.5)
4	Support vector regression	0.203 (0.161-0.244)	0.476 (0.406-0.554)	0.442 (0.418-0.469)
5	Nearest neighbors	0.101 (0.055-0.140)	0.529 (0.460-0.597)	0.502 (0.477-0.528)
6	Decision tree regression	0.154 (0.108-0.198)	0.492 (0.427-0.563)	0.471 (0.447-0.495)
7	Recurrent NN[d]	0.21 (0.165-0.254)	0.459 (0.4-0.523)	0.454 (0.432-0.477)

[a]MSE: mean-squared error.

[b]MAE: mean absolute error.

[c]GLM: generalized linear model.

[d]NN: neural network.

Prediction of aPTT by the Recurrent Neural Network Model

In this section we present the results of the recurrent neural network model and compare predictions with measured values on the test set. Multiple experiments with the model were performed, and the best handpicked parameters are shown in Table 5.

Predictions and measurements are shown in Figure 4. The distributions of aPTT values in the test data alone show a similar distribution as the aPTT values over the entire data set (cf Figure 3 and Figure 4 right panel). The histogram of predictions of the recurrent neural network model has a similar shape (cf Figure 4 top panel and Figure 4 right panel).

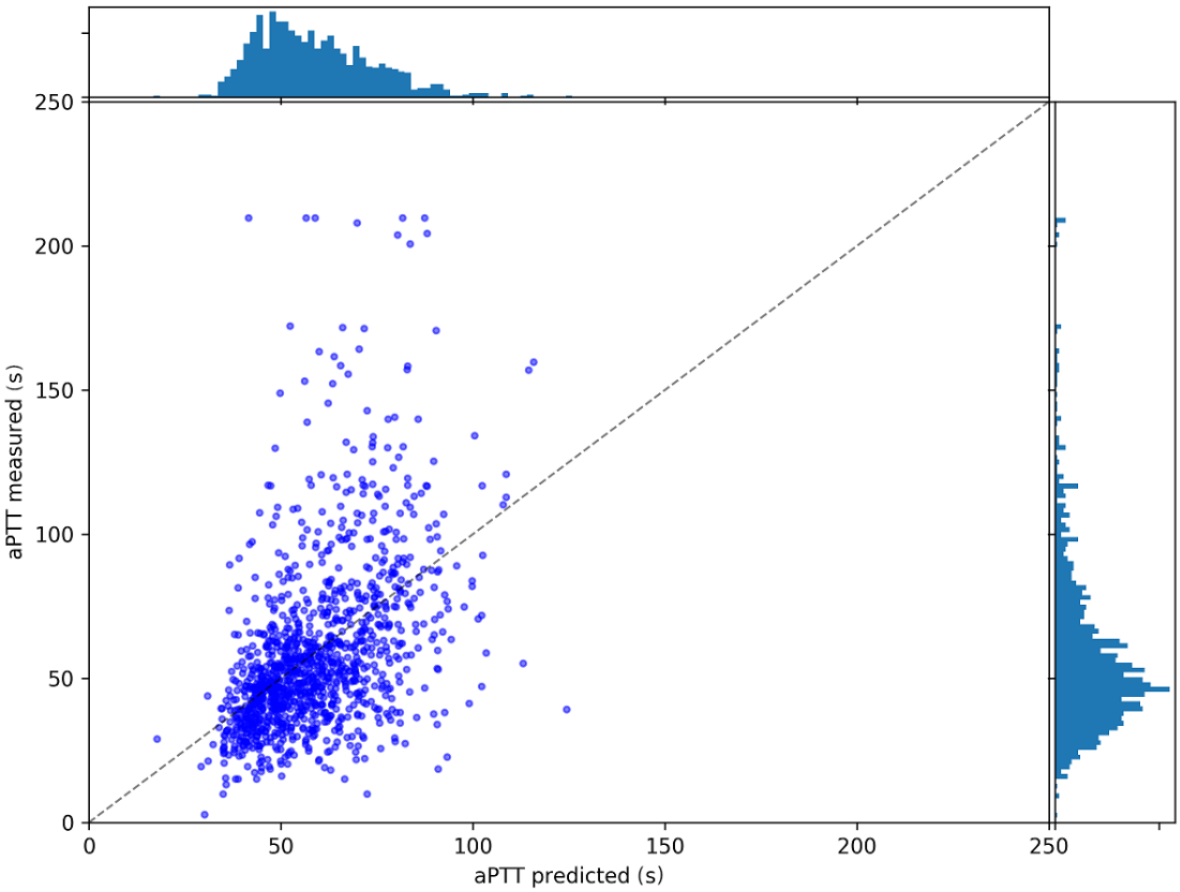
Direct comparisons between predictions and measurements can be seen in the center of Figure 4. The model can predict the majority of aPTT values very well. Although some outliers are predicted accurately, there are a few outliers above 150 s where predictions fall below 75 s. Likewise, some predicted outliers do not manifest as actual outliers.

Table 5. Best hyperparameters for the recurrent neural network model.

Parameter	Value
Learning rate	$1e^{-3}$
Layers	Single GRU[a] layer; 3 feedforward layers with 10, 5, and 1 output neurons, respectively
Hidden size (GRU)	5
Bidirectional	True
Accumulate gradient batches	16
L2 penalty on all weights	0.2

[a]GRU: gated recurrent unit.

Figure 4. Predictions versus measurements. The figure shows predicted (abcissa) and measured aPTT (ordinate) after 24 hours in the central panel. Only predictions on the test set are shown. The dashed diagonal line indicates a perfect match between prediction and measurement. Above and to the right are binned distributions of all predictions and measurements, respectively. aPTT: activated partial thromboplastin time.



Comparison With Classification Models

In the previous sections, we have seen that the recurrent neural network shows the highest performance on the regression task. However, it is also apparent that not all predictions are accurate. To understand whether improvements needed to occur on the models or on data quality aspects, we rephrased the problem as a classification task to be able to compare the performance of the trained model with the 3 most recently published classification models [14,15,18]. Each of the 3 models was trained on our data set (details in the Methods section).

Our recurrent neural network scored the highest performance in recall and F_1-score. The simplest model (logistic regression by Ghassemi et al [14]) had the highest precision, and the feedforward neural network by Li et al [18] had the highest accuracy (see Table 6 for results). No single model outperformed the others on all 4 metrics, and the appropriate model may be chosen depending on which metric is considered most relevant.

The fact that the best-published models show a comparable performance indicate that significant improvements require a closer monitoring of patients, additional tests, and improved data quality.

Table 6. Comparison of different models when formulating activated partial thromboplastin time prediction as a classification task. For each metric, a higher score is better.

Model	Precision	Recall	F_1-score	Accuracy
GRU[a]	0.411	0.396	0.398	0.829
Ghassemi [14]	0.707	0.357	0.356	0.825
Su [15]	0.357	0.338	0.316	0.834
Li [18]	0.430	0.350	0.338	0.838

[a]GRU: gated recurrent unit.

Discussion

Principal Findings

In this study, we analyzed and predicted the effect of heparin treatment on a cohort of 5742 patients and 5926 hospital admissions 24 hours after continuous application. A statistically significant shift of aPTT measurements compared to the beginning of the treatment was observed. Most patients' aPTT measurements were within 35 s to 75 s; however, some patients showed much higher aPTT values, leading to a challenging prediction problem with a long-tailed distribution. We demonstrated that ML models can aid in predicting the aPTT values 12 hours in advance. Additionally, we have shown that using the time series of variables improves predictive performance.

Some underlying medical conditions, while occurring rarely, are known to cause much higher aPTT values. These medical conditions include lupus anticoagulants or deficiencies in the intrinsic (deficiency in factors IX or X) or extrinsic pathways (deficiency in factors VII) [30,31]. These conditions are not routinely checked for and are only diagnosed when advanced lab testing is ordered.

Established guidelines aim for a prolongation of aPTT by 1.5 to 2.5 times [11-13]. Since patients have different aPTT values before heparin is administered, the target value according to the guidelines is different. Furthermore, medical professionals may define individual anticoagulation targets that do not match a prolongation of 1.5 to 2.5 times the baseline value. Thus, we consider aPTT prediction to be a regression problem as Kong et al [16] and Smith et al [17] have done. A model that predicts aPTT several hours before blood is drawn and analyzed can serve as a valuable aid in adjusting the heparin dosing to meet the patient's aPTT target earlier.

In principle, aPTT can be predicted continuously. However, to allow a comparison between models that make a single prediction based on measurements at a single point in time and a model that consumes the entire time series, we fixed 2 time points (at 12 hours and 24 hours after continuous treatment started). Models can use data available at 12 hours and make a prediction for 24 hours after continuous treatment starts. The cutoff after 12 hours is arbitrary and could be reasonably made at a different time. The second point in time is motivated by the observation that reaching the aPTT target within 24 hours is associated with favorable outcomes [6]. The recurrent neural network showed the best performance, and its predictions were analyzed in detail. Although most samples were predicted well, an unsolved problem is that rare cases exhibit a remarkably high aPTT and are not captured by the model. As mentioned earlier, underlying medical conditions are known to cause significantly longer aPTT. We hypothesized that, for significantly improved predictions, either testing of conditions that cause a long aPTT or much more frequent measurements of aPTT combined with dosing adjustments are required.

Recent literature on aPTT prediction after heparin treatment considers 3 distinct ranges [14,15,18]. In order to compare our model to those in the literature, we binned our predictions into subtherapeutic, therapeutic, and supratherapeutic as introduced by Ghassemi et al [14]. We observed that our model showed a higher recall and F_1-score than did the other models. Arguably, the setup that we chose was the most difficult compared to the references since we predicted a single aPTT value 12 to 36 hours in advance. Others made predictions 4 to 6 hours [15] or 4 to 8 hours [14] in advance or averaged aPTT measurements between 4 and 24 hours [18].

Limitations

Other anticoagulants, such as warfarin or argatroban, were not considered. We expect that only a small sample of patients, if any, are receiving heparin together with anticoagulants and, therefore, decided not to take it into account as is common in similar studies [19].

It is well known that the laboratory conditions can affect the ranges of aPTT measurements [32]. The aPTT measurements were all reported by the same laboratory. Thus, the model may not be applicable to other centers and laboratories without parameter fine-tuning.

Modeling decisions that may negatively affect the model performance are the resampling of time series to 2-hour intervals. This resampling might miss significant changes in some variables. Furthermore, handling of missing data by forward and mean imputation could be improved by multiple imputation methods.

Conclusions

Anticoagulation therapy with heparin monitored by the aPTT laboratory assay is a widely used procedure in ICUs. It is well known that heparin dosing is challenging due to high interpatient variability. In the future, ML may help to suggest personalized dosing recommendations. We demonstrated that a model based on time series performs best.

Acknowledgments

We acknowledge financial support from the Open Access Publication Fund of Charité – Universitätsmedizin Berlin and the German Research Foundation (DFG).

Conflicts of Interest

FB reports grants from the German Federal Ministry of Education and Research, the German Federal Ministry of Health, the Berlin Institute of Health, the Hans Böckler Foundation, the Einstein Foundation, the Berlin University Alliance, and the Robert Koch Institute; and personal fees from Elsevier Publishing, Medtronic, and GE Healthcare outside the submitted work.

Multimedia Appendix 1
The best hyperparameters of the static models along with additional evaluation metrics.
[DOC File , 93 KB - medinform_v10i10e39187_app1.doc]

References

1. Martin KA, Molsberry R, Cuttica MJ, Desai KR, Schimmel DR, Khan SS. Time trends in pulmonary embolism mortality rates in the United States, 1999 to 2018. J Am Heart Assoc 2020 Sep;9(17):e016784 [FREE Full text] [doi: 10.1161/JAHA.120.016784] [Medline: 32809909]
2. Hald EM, Løchen ML, Mathiesen EB, Wilsgaard T, Njølstad I, Brækkan SK, et al. Atrial fibrillation, venous thromboembolism, ischemic stroke, and all-cause mortality: The Tromsø study. Res Pract Thromb Haemost 2020 Aug;4(6):1004-1012 [FREE Full text] [doi: 10.1002/rth2.12351] [Medline: 32864551]
3. Heit JA, Mohr DN, Silverstein MD, Petterson TM, O'Fallon WM, Melton LJ. Predictors of recurrence after deep vein thrombosis and pulmonary embolism: a population-based cohort study. Arch Intern Med 2000 Mar 27;160(6):761-768. [doi: 10.1001/archinte.160.6.761] [Medline: 10737275]
4. Wang H, Cushman M, Rosendaal FR, van Hylckama Vlieg A. Association of remote history of venous thrombosis with risk of venous thrombosis after age 70 years. JAMA Netw Open 2022 Mar 01;5(3):e224205 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.4205] [Medline: 35333359]
5. Bartlett MA, Mauck KF, Stephenson CR, Ganesh R, Daniels PR. Perioperative venous thromboembolism prophylaxis. Mayo Clin Proc 2020 Dec;95(12):2775-2798. [doi: 10.1016/j.mayocp.2020.06.015] [Medline: 33276846]
6. Smith SB, Geske JB, Maguire JM, Zane NA, Carter RE, Morgenthaler TI. Early anticoagulation is associated with reduced mortality for acute pulmonary embolism. Chest 2010 Jun;137(6):1382-1390 [FREE Full text] [doi: 10.1378/chest.09-0959] [Medline: 20081101]
7. Minet C, Potton L, Bonadona A, Hamidfar-Roy R, Somohano CA, Lugosi M, et al. Venous thromboembolism in the ICU: main characteristics, diagnosis and thromboprophylaxis. Crit Care 2015 Aug 18;19:287 [FREE Full text] [doi: 10.1186/s13054-015-1003-9] [Medline: 26283414]
8. Konstantinides S, Meyer G, Becattini C, Bueno H, Geersing G, Harjola V, The Task Force for the diagnosismanagement of acute pulmonary embolism of the European Society of Cardiology (ESC). 2019 ESC Guidelines for the diagnosis and management of acute pulmonary embolism developed in collaboration with the European Respiratory Society (ERS): The Task Force for the diagnosis and management of acute pulmonary embolism of the European Society of Cardiology (ESC). Eur Respir J 2019 Sep;54(3):1901647 [FREE Full text] [doi: 10.1183/13993003.01647-2019] [Medline: 31473594]
9. Konstantinides S, Torbicki A, Agnelli G, Danchin N, Fitzmaurice D, Galiè N, et al. 2014 ESC Guidelines on the diagnosis and management of acute pulmonary embolism. Kardiol Pol 2014 Nov 14;72(11):997-1053. [doi: 10.5603/kp.2014.0211]
10. Levine MN, Raskob G, Landefeld S, Hirsh J. Hemorrhagic complications of anticoagulant treatment. Chest 1995 Oct;108(4 Suppl):276S-290S. [doi: 10.1378/chest.108.4_supplement.276s] [Medline: 7555182]
11. Garcia DA, Baglin TP, Weitz JI, Samama MM. Parenteral anticoagulants: antithrombotic therapy and prevention of thrombosis, 9th ed: American College of Chest Physicians evidence-based clinical practice guidelines. Chest 2012 Feb;141(2 Suppl):e24S-e43S [FREE Full text] [doi: 10.1378/chest.11-2291] [Medline: 22315264]
12. Basu D, Gallus A, Hirsh J, Cade J. A prospective study of the value of monitoring heparin treatment with the activated partial thromboplastin time. N Engl J Med 1972 Aug 17;287(7):324-327. [doi: 10.1056/nejm197208172870703]
13. Eikelboom J, Hirsh J. Monitoring unfractionated heparin with the aPTT: time for a fresh look. Thromb Haemost 2017 Dec 01;96(11):547-552. [doi: 10.1160/th06-05-0290]
14. Ghassemi MM, Richter SE, Eche IM, Chen TW, Danziger J, Celi LA. A data-driven approach to optimized medication dosing: a focus on heparin. Intensive Care Med 2014 Sep;40(9):1332-1339 [FREE Full text] [doi: 10.1007/s00134-014-3406-5] [Medline: 25091788]
15. Su L, Liu C, Li D, He J, Zheng F, Jiang H, et al. Toward optimal heparin dosing by comparing multiple machine learning methods: retrospective study. JMIR Med Inform 2020 Jun 22;8(6):e17648 [FREE Full text] [doi: 10.2196/17648] [Medline: 32568089]
16. Kong N, Liu X, Liu C, Lian J, Wang H. Deep architecture for Heparin dosage prediction during continuous renal replacement therapy. : IEEE; 2017 Presented at: 36th Chinese Control Conference (CCC); July 26-28, 2017; Dalian, China URL: http://ieeexplore.ieee.org/document/8029139/ [doi: 10.23919/chicc.2017.8029139]
17. Smith BP, Ward RA, Brier ME. Prediction of anticoagulation during hemodialysis by population kinetics and an artificial neural network. Artif Organs 1998 Sep;22(9):731-739. [doi: 10.1046/j.1525-1594.1998.06101.x] [Medline: 9754457]
18. Li D, Gao J, Hong N, Wang H, Su L, Liu C, et al. A clinical prediction model to predict heparin treatment outcomes and provide dosage recommendations: development and validation study. J Med Internet Res 2021 May 20;23(5):e27118 [FREE Full text] [doi: 10.2196/27118] [Medline: 34014171]
19. Falconer N, Abdel-Hafez A, Scott IA, Marxen S, Canaris S, Barras M. Systematic review of machine learning models for personalised dosing of heparin. Br J Clin Pharmacol 2021 Nov;87(11):4124-4139. [doi: 10.1111/bcp.14852] [Medline: 33835524]

20. Salgado C, Azevedo C, Proença H, Vieira S. Missing data. In: Secondary Analysis of Electronic Health Records Internet. New York, NY: Springer International Publishing; Sep 10, 2016:143-162.
21. Vollmer S, Mateen BA, Bohner G, Király FJ, Ghani R, Jonsson P, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. BMJ 2020 Mar 20;368:l6927 [FREE Full text] [doi: 10.1136/bmj.l6927] [Medline: 32198138]
22. Li J, Yan XS, Chaudhary D, Avula V, Mudiganti S, Husby H, et al. Imputation of missing values for electronic health record laboratory data. NPJ Digit Med 2021 Oct 11;4(1):147 [FREE Full text] [doi: 10.1038/s41746-021-00518-0] [Medline: 34635760]
23. Che Z, Purushotham S, Cho K, Sontag D, Liu Y. Recurrent neural networks for multivariate time series with missing values. Sci Rep 2018 Apr 17;8(1):6085 [FREE Full text] [doi: 10.1038/s41598-018-24271-9] [Medline: 29666385]
24. Abraham A, Pedregosa F, Eickenberg M, Gervais P, Mueller A, Kossaifi J, et al. Machine learning for neuroimaging with scikit-learn. Front Neuroinform 2014;8:14 [FREE Full text] [doi: 10.3389/fninf.2014.00014] [Medline: 24600388]
25. Nair V, Hinton G. Rectified linear units improve restricted boltzmann machines. 2010 Jun 24 Presented at: International Conference on Machine Learning; June 21 - 24, 2010; Haifa, Israel.
26. Kingma D, Ba J. Adam: A method for stochastic optimization. ArXiv Prepr ArXiv14126980. 2014. URL: https://arxiv.org/pdf/1412.6980.pdf [accessed 2022-05-01]
27. aptt-prediction. GitHub. URL: https://github.com/sebboie/aptt-prediction [accessed 2022-05-17]
28. Botev Z, Kroese D, Rubinstein R, L'Ecuyer P. The cross-entropy method for optimization. In: Handbook of Statistics Internet. Amsterdam, the Netherlands: Elsevier; 2013.
29. Grandini M, Bagli E, Visani G. Metrics for multi-class classification: an overview. arXiv. Metrics for Multi-Class Classification URL: http://arxiv.org/abs/2008.05756 [accessed 2022-07-15]
30. Chng W, Sum C, Kuperan P. Causes of isolated prolonged activated partial thromboplastin time in an acute care general hospital. Singapore Med J 2005 Sep;46(9):450-456 [FREE Full text] [Medline: 16123828]
31. Barbosa ACN, Montalvão SAL, Barbosa KGN, Colella MP, Annichino-Bizzacchi JM, Ozelo MC, et al. Prolonged APTT of unknown etiology: A systematic evaluation of causes and laboratory resource use in an outpatient hemostasis academic unit. Res Pract Thromb Haemost 2019 Oct 08;3(4):749-757 [FREE Full text] [doi: 10.1002/rth2.12252] [Medline: 31624795]
32. Toulon P, Smahi M, De Pooter N. APTT therapeutic range for monitoring unfractionated heparin therapy. Significant impact of the anti-Xa reagent used for correlation. J Thromb Haemost 2021 Aug;19(8):2002-2006. [doi: 10.1111/jth.15264] [Medline: 33555096]

Abbreviations

APACHE II: acute physiology and chronic health evaluation II
aPTT: activated partial thromboplastin time
Charité: Charité – Universitätsmedizin Berlin
ECMO: extracorporeal membrane oxygenation
GRU: gated recurrent unit
ICU: intensive care unit
MIMIC: Multiparameter Intelligent Monitoring in Intensive Care
ML: machine learning
MSE: mean-squared error
SAPS II: simplified acute physiology score II
SOFA: sequential organ failure assessment
SVR: support vector machine regression
TISS-10: therapeutic intervention scoring system

Edited by C Lovis, J Hefner; submitted 02.05.22; peer-reviewed by H Zhang, F Chen; comments to author 24.05.22; revised version received 17.07.22; accepted 11.08.22; published 13.10.22.

Please cite as:
Boie SD, Engelhardt LJ, Coenen N, Giesa N, Rubarth K, Menk M, Balzer F
A Recurrent Neural Network Model for Predicting Activated Partial Thromboplastin Time After Treatment With Heparin: Retrospective Study
JMIR Med Inform 2022;10(10):e39187
URL: https://medinform.jmir.org/2022/10/e39187
doi:10.2196/39187
PMID:36227653

Original Paper

Relation Extraction in Biomedical Texts Based on Multi-Head Attention Model With Syntactic Dependency Feature: Modeling Study

Yongbin Li[1], ME; Linhu Hui[1], ME; Liping Zou[1], ME; Huyang Li[1], ME; Luo Xu[1], PhD; Xiaohua Wang[1], PhD; Stephanie Chua[2], PhD

[1]School of Medical Information Engineering, Zunyi Medical University, Zunyi, China

[2]Faculty of Computer Science and Information Technology, University Malaysia Sarawak, Sarawak, Malaysia

Corresponding Author:
Yongbin Li, ME
School of Medical Information Engineering
Zunyi Medical University
6 Xuefu Road West, Xinpu New District
Zunyi, 563000
China
Phone: 86 18311545098
Email: bynn456@126.com

Abstract

Background: With the rapid expansion of biomedical literature, biomedical information extraction has attracted increasing attention from researchers. In particular, relation extraction between 2 entities is a long-term research topic.

Objective: This study aimed to perform 2 multiclass relation extraction tasks of Biomedical Natural Language Processing Workshop 2019 Open Shared Tasks: relation extraction of Bacteria-Biotope (BB-rel) task and binary relation extraction of plant seed development (SeeDev-binary) task. In essence, these 2 tasks are aimed at extracting the relation between annotated entity pairs from biomedical texts, which is a challenging problem.

Methods: Traditional research methods adopted feature- or kernel-based methods and achieved good performance. For these tasks, we propose a deep learning model based on a combination of several distributed features, such as domain-specific word embedding, part-of-speech embedding, entity-type embedding, distance embedding, and position embedding. The multi-head attention mechanism is used to extract the global semantic features of an entire sentence. Meanwhile, we introduced a dependency-type feature and the shortest dependency path connecting 2 candidate entities in the syntactic dependency graph to enrich the feature representation.

Results: Experiments show that our proposed model has excellent performance in biomedical relation extraction, achieving F_1 scores of 65.56% and 38.04% on the test sets of the BB-rel and SeeDev-binary tasks. Especially in the SeeDev-binary task, the F_1 score of our model is superior to that of other existing models and achieves state-of-the-art performance.

Conclusions: We demonstrated that the multi-head attention mechanism can learn relevant syntactic and semantic features in different representation subspaces and different positions to extract comprehensive feature representation. Moreover, syntactic dependency features can improve the performance of the model by learning dependency relation between the entities in biomedical texts.

(JMIR Med Inform 2022;10(10):e41136) doi:10.2196/41136

KEYWORDS

biomedical relation extraction; deep learning; feature combination; multi-head attention; additive attention; syntactic dependency feature; syntactic dependency graph; shortest dependency path

Introduction

Background

Information extraction (IE) [1] involves extracting specific events or related information from texts; automatically classifying, extracting, and reconstructing useful information from massive amounts of content; and transforming it into structured knowledge. With the increasing demand for text mining technology to locate key information in biomedical literature, biomedical IE [2,3] has become a new research hot spot. Simultaneously, with the explosive development of biomedical literature, many research directions for biomedical IE have been promoted, such as named entity recognition, protein relation extraction [4], and drug interaction extraction [5]. In particular, it is a challenging and practical problem to detect the relation between annotated entities in the biomedical text under relation constraints, which is an important research direction.

The Biomedical Natural Language Processing Workshop-Open Shared Task (BioNLP-OST) series [6] is representative of biomolecular IE, which aims to facilitate the development and sharing of biomedical text mining and fine-grained IE. BioNLP-OST has made a great contribution to the development of biomedical IE and has been held for 5 times. The research topics of BioNLP-OST include fine-grained event extraction, biomedical knowledge base construction, and other scopes. This study mainly focused on the relation extraction of Bacteria-Biotope (BB-rel) task and the binary relation extraction of plant seed development (SeeDev-binary) task in BioNLP-OST 2019 [7]. These 2 multiclass subtasks are essential for predicting whether and what relationship exists between 2 annotated entities. This study contributes to the development of practical applications for biomedical text mining.

A series of innovative systems have achieved good results and actively promoted the development of biomedical IE. For example, in BB-rel and SeeDev-binary tasks, traditional relation extraction models are mainly based on feature-based [8,9] and kernel-based methods [10,11]. These methods rely on domain-specific knowledge or language tools to extract artificial features. For example, in the study by Björne and Salakoski [12], a relation extraction system was constructed using a feature based on the shortest dependent path and support vector machine (SVM). In recent years, deep learning (DL) models have been successfully applied in many fields of natural language processing, requiring less feature engineering and automatic learning of useful information from corpus data (Kumar, S, unpublished data, May 2017). In the biomedical relation extraction field, several well-known DL models have been gradually applied and have achieved excellent performance, including distributed representation [13,14], convolutional neural network (CNN) [15-17], and recurrent neural network [18-20]. Consequently, instead of complicating handcrafted feature engineering, we used the DL method to extract relations in biomedical texts.

The combined application of the distributed features of a full sentence is the most common method for biomedical relation extraction [13,21,22]. Here, we use a variety of distributed features, such as domain-specific word embedding [23], part of speech (POS) embedding [24], entity-type embedding [13], and distance embedding [25]. However, the commonly used model is difficult to focus on the key information of full sentence; therefore, the attention mechanism [26] has been proposed and proven to be successful in a wide range of natural languages processing fields, such as machine translation, reading comprehension, and sentiment classification [27-29]. In our proposed model, we use the multi-head attention mechanism proposed by Vaswani et al [30] to deal with the combination of distributed features of the full sentence. Multi-head attention can ignore the distance between words, directly calculate the dependency between words, and learn the syntactic and semantic features of sentences in different representation subspaces. We also constructed position embedding (PE) to inject position information to take advantage of the order of words in a sentence.

In our proposed model, we also integrated the shortest dependency path and dependency-type feature based on the syntactic dependency graph as one of the input features, which has been proven to be effective in several studies [19,31,32]. Although syntactic dependency features contain valuable syntactic information to facilitate the extraction of biomedical relations, they may still lose important information, such as prepositions before or after entities are likely to be discarded on the dependency path, which should play a key role [33]. Hence, this study adopts the combination of distributed features and syntactic dependency features as the final feature representation of biomedical texts, in which syntactic dependency features exist as supplementary features.

In this paper, we introduce a DL model to solve 2 biomedical relation extraction tasks: SeeDev-binary and BB-rel. We combined several distributed features and a multi-head attention mechanism to automatically extract global semantic features from long and complicated sentences. Syntactic-dependent features were also integrated into the model. As the shortest dependency path connecting 2 entities is short and concise, we apply a CNN to learn its features. We conducted extensive experiments, and our approach achieved F_1 scores of 65.56% and 38.04% on BB-rel and SeeDev-binary tasks and achieved state-of-the-art performance on the SeeDev-binary task.

Related Work

The BB-rel task was conducted 3 times [34] before, and the fourth edition [35] in the BioNLP-OST 2019 focused on extracting information about bacterial biotopes and phenotypes, motivated by the importance of knowledge on biodiversity for theoretical research and applications in microbiology, involving entity recognition, entity normalization, and relation extraction. This edition has been extended to include a new entity type of *phenotype*, relation category of *Exhibits*, and new documents. We mainly studied one of the subtasks, the relation extraction task (BB-rel), which is to predict the relationship of *Lives_In* category between *microorganisms*, *habitats*, and *geographic* entities, and the relation of *Exhibits* category between *microorganism* and *phenotype* entities from PubMed abstracts and full-text excerpts, where entity annotation has been provided. Many researchers have contributed their efforts to the

BB-rel task and have proposed innovative methods. For example, in Biomedical Natural Language Processing Workshop 2016, TurkuNLP team used the method of the shortest dependent path using the Turku event extraction system (TEES) [12] and 3 long short-term memory (LSTM) units, achieving an F_1 score of 52.10% [31]. The bidirectional gated recurrent unit-Attn team proposed a bidirectional gated recurrent unit with an attention model, with an F_1 score of 57.42% [36]. Amarin et al [33] combined feature combinations with an attention model and contextual representations to achieve a state-of-the-art performance with an F_1 score of 60.77%. In BioNLP-OST 2019, almost all researchers used neural network models in various architectures. For instance, the Yuhang_Wu team used a multilayer perceptron and achieved an F_1 score of 60.49% on the test set. The highest F_1 score was 66.39%, which was submitted by the whunlp team [37]. They constructed a dependency graph based on lexical association, and used bidirectional LSTM (BiLSTM) [38] and an attention graph convolution neural network to detect the relation. In addition, the AliAI team innovatively used a multitask architecture similar to *Bidirectional Encoder Representations from Transformers* (BERT) and achieved 64.96%, which effectively alleviated the lack of information in the domain-specific field [39].

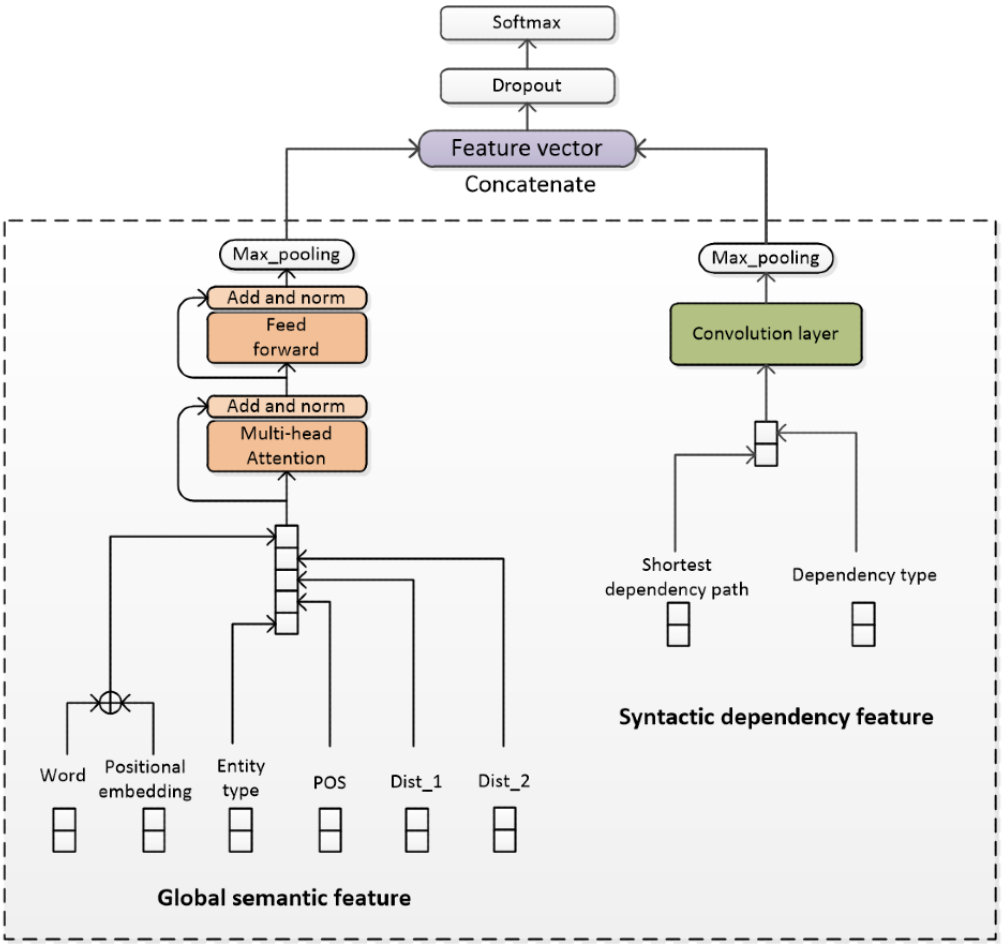
The SeeDev task [40] aims to facilitate the extraction of complex events on regulations in plant development from scientific articles, with a focus on events describing the genetic and molecular mechanisms involved in *Arabidopsis thaliana* seed development. The SeeDev task involves extracting 21 relation categories, involving 16 entity types, to accurately reflect the complexity of the regulatory mechanisms of seed development, which is a major scientific challenge. SeeDev was originally proposed at BioNLP-OST 2016 [6], and in 2019, the evaluation methodology focused more on the contribution of biology. It includes full and binary relation extraction, in which we mainly study the binary relation extraction subtask SeeDev-binary. To address this problem, most researchers have used traditional supervised machine learning approaches. These systems design artificial templates or manually extract many features based on domain-specific knowledge, such as linguistic features, semantic features, and syntactic information, which are added to the system as feature representations. Kernel-based machine learning algorithms such as SVM and Bayesian are then used to detect the relation categories, which are widely used for IE. For instance, the UniMelb team [41] developed an event extraction system using rich feature sets and SVM classifiers with a linear kernel. In addition, the MIC-CIS team [42] used an SVM combined with linguistic features to achieve optimal results on BioNLP-OST 2019. As the DL model gradually became the main research method, the DUTIR team [13] innovatively used a DL model based on distributed features and a CNN model [15]. The YNU-junyi team [14] integrated the LSTM model [18] based on a CNN model to address the problem that CNN alone cannot capture the long-range dependence of sequences, and they obtained an F_1 score of 34.18% on the SeeDev-binary task of BioNLP-OST 2019.

Methods

Overview

In this section, we describe our proposed model for the 2 biomedical relation extraction tasks in detail. The overall architecture is shown in Figure 1. The preprocessing of the data sets is described in the first part. In the second part, we introduce a series of distributed semantic features used in our method, and the multi-head attention mechanism used on them is introduced in the third part. The fourth part explains the construction of the syntactic dependency feature. In the fifth part, we introduce the classification and training details. Finally, we present the training and hyperparameter settings.

Figure 1. The overall architecture of our proposed model with global semantic feature based on feature combination and multi-head attention as well as syntactic dependency feature. Dist_1: distance embedding corresponding to the first entity in a sentence; Dist_2: distance embedding corresponding to the second entity in a sentence; entity type: entity type embedding; POS: part-of-speech embedding; Word: word embedding.



Data Preprocessing

In the data preprocessing phase, we used TEES [12,31] to run a text preprocessing pipeline. The TEES system splits the text into sentences using the GENIA Sentence Splitter [43] and parses the sentences through the integrated the Brown Laboratory for Linguistic Information Processing parser [44] with the biomedical domain model [45] to obtain the tokens, POS tags, and parse graphs for each word. Then, the phrase structure trees obtained by the parser are further processed using the Stanford conversion tool [46] to obtain the syntactic dependency graph.

The BB-rel and SeeDev-binary tasks are relation extraction tasks, which detect whether and what relations exist between 2 annotated entities in biomedical texts. For example, in the sentence "The percentage of penicillin-resistant *N. gonorrhoeae* isolated in the region over the decade varied considerably," in which *N. gonorrhoeae* is a microorganism-type entity and "percentage" is a phenotype-type entity, we need to detect whether there is a relationship between them and the category of the relation. There are usually 2 solutions to the relation extraction task: the first is to identify whether there is a relation between entity pairs in a sentence and then classify a correct category [47], and the second method is to combine the 2 steps of identification and classification into 1 step [13]. This paper adopts the second method, which regards nonrelation as a category of relationships and carries out multi-category classification.

In the training and validation sets of the BB-rel and SeeDev-binary tasks, only positive instances were labeled. However, in the prediction phase, there may be a nonrelation between 2 candidate entities; therefore, it is necessary to manually construct negative instances in the training phase. After the biomedical texts are divided into sentences, we enumerate each entity pair in the sentence and judge the unlabeled instances as nonrelational. Because the biomedical relation extraction of SeeDev-binary and BB-rel tasks is under the constraint of regulation, there must be no relation between some entity types. For example, in the BB-rel task, there must be no biomedical relation between the entity of *geographic* type and the entity of *phenotype* type. Therefore, we need to further eliminate the entity pairs that do not comply with the regulations.

In the data sets of the 2 tasks, not only do the entities of a relation appear in the same sentence (intrasentence) but also the entities of a relation may be in different sentences (intersentence), which is a great challenge regarding biomedical relation extraction tasks [35]. In our method, we only considered intrasentence relations and ignored intersentence relations. There are 2 difficulties involved in the intersentence relation: one is that the reasoning relationship is difficult and complex; the other is that the number of negative instances increases exponentially,

which leads to an extreme imbalance of positive and negative samples, resulting in performance degradation of the model. Therefore, all existing systems only extract intrasentence relations without considering intersentence relations [35,40]. In addition, an instance is eliminated if there is no syntactic dependency path between the 2 candidate entities.

Distributed Semantic Representation

Our method extracts global semantic features from a full sentence through a combination of several distributed features and a multi-head attention mechanism. Domain-specific word embedding, POS embedding, entity-type embedding, distance embedding, and PE were integrated into our model.

Word embedding is a frequently used distributed representation model that encodes rich semantic information into vectors. The sequence of a full sentence of length n can be represented as $\{w_1,e_1,...,e_2,w_n\}$, where e_1 and e_2 represent entity pairs. We initialized our word embeddings with a pretrained 200-dimensional biomedical word embedding model [23], which was trained on PubMed and PMC abstracts, and full texts contained an unannotated corpus of 5 billion tokens. The pretrained embedding model was trained using the word2vec tool with the skip-gram model [48]. We only used the most frequent 100k words to build dictionary D, and the unknown words in the data sets were randomly initialized. Taking the BB-rel task as an example, it is possible that the words of entity are not in dictionary D, so we add the words "Microorganism," "Habitat," "Geographical," and "phenotype" to the dictionary and initialize them randomly. If an entity is of *microorganism* type and is not in the word embedding model, it will be replaced by the word "Microorganism." Through the pretrained word embedding matrix, we can transform the sequence of tokens in a full sentence into a vector sequence  . We also used POS embedding [24] to encode the POS for words in a sentence, which usually plays an important role. The POS embedding was randomly initialized and fine-tuned during the training phase.

The combination of different types of entities has different probabilities for some relations; therefore, the entity type is an important factor for prediction [13]. As the 2 biomedical relation extraction tasks are conditionally constrained, they do not involve the direction between entity pairs, so the entity-type sequence only needs one chain to represent. Therefore, the entity-type sequence can be expressed as $\{-1,t_1,...,t_2,-1\}$, where nonentity words are labeled as −1. Through a randomly initialized type embedding matrix, the entity-type vector sequence can be represented as  .

The distance sequence is divided into 2 chains, namely, the distance from the current word to the 2 candidate entities. In our method, relative distance [25] is used to measure the distance between the current word and an entity, which can be formulated as equation 1, where l is the absolute distance and s is the maximum distance in the data sets. As the relative distance is not an integer, it is necessary to construct a distance dictionary and use the distance embedding matrix to generate the distance-vector sequence.

As we use the multi-head attention model to deal with the combination of a series of distributed features without using any time series model, we have to inject some absolute position information of words into the model; therefore, we introduce PE with reference as shown in the study by Vaswani et al [30]. In our method, the PE vectors have the same dimension d_{word} as the word embedding, and then PE vectors can be calculated according to the sine and cosine functions of the frequencies. The formulas are given in equations 2 and 3, where *pos* is the position and i represents the i-th dimension of one word. Finally, the position information was injected into the model by adding the PE vector into the word embedding.

Finally, a series of distributed features is concatenated, and each word w_i in the sentence can be represented as  . This comprehensive distributed feature is sent to the multi-head attention layer to extract the global semantic features of the full sentence.

Multi-Head Attention Layer

In recent years, a series of attention-based models have been applied to relationship extraction tasks with remarkable success [49,50]. The core idea of the attention mechanism is to locate key information from text by assigning attention scores. At present, the most widely used attention models are additive attention [26] and dot-product attention [30]. In the study by Vaswani et al [30], the multi-head attention mechanism was proposed as the main component unit of the transformer model. In this model, attention can be used to compute the output of a series of values through value mapping to a set of key-value pairs, that is, to calculate a weighted sum of the values, where the weight assigned to each value is computed by a query with the corresponding key. In our method, the multi-head attention mechanism is used as an encoder to extract the global semantic feature of the full sentence, and each attention head is calculated by integrating the position information and using the scaled dot-product attention function.

The overall structure of scaled dot-product attention and multi-head attention is shown in Figure 2, similar to that shown in the study by Vaswani et al [30]. Here, Q, K, and V are the same, which are the feature combinations from the full sentence; therefore, multi-head attention can also be understood as a form of self-attention. Eight attention heads based on scaled dot-product attention were used to extract features, which divided feature combinations into 8 channels. For each channel, the embedding of each word in the sentence with length n can be expressed as z_i. Through the weights (W_q, W_k, W_v) that are not shared between channels, we can get the vector expression of a word in different subspaces, namely (q_i, k_i, v_i), as shown in equation 4.

The attention weight vector a_i corresponding to i-th query is calculated by the dot product of the query vector and key vector and then scaled by  and calculated by a Softmax function, where d^k is the dimensionality of the feature combination and n is the length of the sentence, as shown in equation 5.

By multiplying the attention weight vector a_i by the value sequence of length n, a feature vector c_i is obtained, which is a weighted sum of the values, as shown in equation 6.

Therefore, the attention head of each channel is a concatenated matrix of n feature vectors, which can be expressed as h_i using equation 7. Each attention head can encode the semantic information of a sentence in subspaces with different representations.

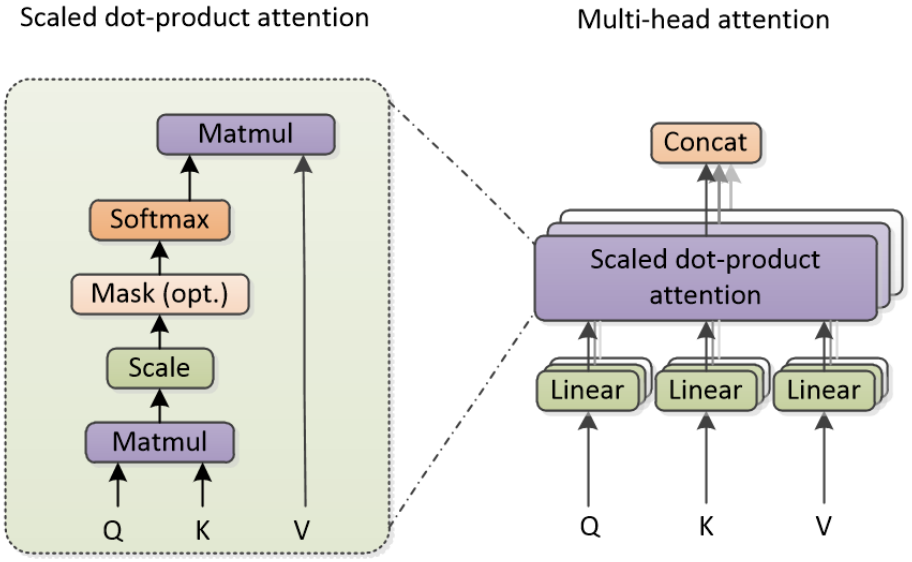
$$h_i = [c_1;c_2;...;c_n] \quad \textbf{(7)}$$

Furthermore, we concatenated multiple attention heads in the last dimension to obtain the multi-head attention feature of the full sentence, as shown in equation 8.

$$MultiHead = [h_1;h_2;...;h_8] \quad \textbf{(8)}$$

Similar to the transformer model, we also used a fully connected neural network behind the multi-head attention model and used a residual join, as shown in Figure 1. Finally, the global semantic features of the full sentence are obtained using a max-pooling operation.

Figure 2. Scaled dot-product attention function (left). Multi-head attention consists of several scaled dot-product attention (right). Concat: concatenate; K: key; Matmul: matrix multiply; Q: query; V: value.



Syntactic Dependency Feature

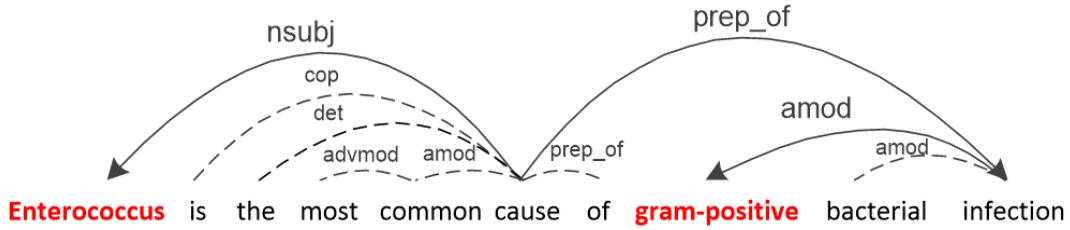
The syntactic dependency features for the proposed DL model are generated based on the shortest dependency path connecting 2 candidate entities and the dependency type in the dependency graph. The shortest dependency path contains the most important terms related to characterizing the extraction and has been successfully applied in relation extraction many times [51,52]. An example of syntactic dependency is shown in Figure 3, where "Enterococcus" is a *microorganism-type* entity and "Gram-positive" is a *phenotype*-type entity. We can observe that the dependency parse between the words is directional. To simplify the calculation, we use the method by Mehryary et al [31] to convert the dependency relation of a sentence into an undirected graph and then find the shortest path between 2 candidate entities using the Dijkstra algorithm. In the case of BB-rel task, we always process from *a microorganism*-type entity to location entities (either a *habitat* or a *geographic* entity) or *phenotype* entity, regardless of their positions in sentences. Therefore, in the example in Figure 3, the shortest dependency path sequence is ("Enterococcus," "cause," "infection," "Gram-positive") and the dependency-type sequence is (nsubj, prep_of, amod).

In this case, the sequence of the shortest dependency path with m tokens can be represented as $\{e_1,w_2,...,e_2\}$, where e_1 and e_2 represent the entity pairs at the head and end of the sequence, respectively. We used the previously mentioned pretrained 200-dimensional biomedical word embedding model [23]. Using the pretrained word embedding model, we can transform the dependent path sequence into a vector sequence  . For the dependent-type sequence $\{t_1,t_2,...,t_{m-1}\}$, we transform it into by randomly initializing the embedding matrix and filling it to the same length as the dependency path. The 2 vector sequences are concatenated, and i-th word can be denoted as  .

To learn the local features of syntactic dependency from the dependency path and dependency type, LSTM [53] are the most frequently used DL models. By observing the length of the shortest dependency path, it is found that most of the interentity dependency lengths are 2 to 5, which belongs to the feature extraction of super-short sequences. Compared with LSTM,

CNN is more suitable for super-short and concise sequences (Yin, W, unpublished data, February 2017). In addition, CNN are more suitable for parallel computing. Hence, we introduced a multifilter CNN model [54] and a max-pooling operation to learn syntactic dependency features, which has the advantage of learning hidden and advanced features from sentences with multiple channels.

Figure 3. An example of syntactic dependency between phenotype-type entity "Enterococcus" and phenotype-type entity "Gram-positive"; solid lines are entity dependencies, and dashed lines are irrelevant dependencies. advmod: adverbial modifier; amod: adjectival modifier; cop: copula; det: determiner; nsubj: nominal subject; prep_of: preposition of.



Classification and Training

In the output layer, we concatenate the global semantic feature vector and syntactic-dependent feature vector of the sentence to obtain a high-quality feature representation of the instance. Furthermore, the dropout algorithm [55] is used to prevent overfitting, the Softmax function is used to classify biomedical relations, and the probability distribution over each relation category is obtained.

The 2 tasks included a training set, validation set, and test set. In the training phase, taking the multi-classification cross entropy as the objective function, the Adaptive moment estimation optimization algorithm [56] with a learning rate of 0.001 was used to update the neural network parameters. The training times determine the generalization performance of the model; that is, too few training epochs lead to underfitting, and overtraining leads to overfitting. Therefore, the traditional early stopping method is adopted in our method, that is, training is stopped when the performance on the validation set is no longer improved. The experimental results show that the training epoch number is not a fixed value and that the model generally converges in approximately 4 epochs.

The data sets of the 2 biomedical relation extraction tasks were relatively small, and the DL model had more training parameters. Consequently, the initial random state of the model may have a significant impact on the final performance of the model, which was verified by a pre-experiment. To reduce the impact of the initialization state on the model, 10 different random initializations were used to evaluate the model, which was to train the same model structure with different random seeds. Finally, the model with the best F_1 score on the validation set was used as the final model. We used the final model to predict the test set and used the results to evaluate our model on a web-based evaluation service.

Parameter Settings

Through the pre-experiment and evaluation based on the validation set, the hyperparameters of our model were determined. The dimensions of domain-specific word embedding, POS embedding, entity-type embedding, distance embedding, PE, and dependency-type embedding were 200, 200, 200, 100, 200, and 200, respectively, and the embedding matrix was fine-tuned during the training phase. For the multi-head attention mechanism, we adopted a single-layer multi-head attention model, in which 8 parallel attention heads were used, and the number of units in the linear layer of each attention head was the same as the input. To extract the syntactic dependency feature, the number of convolution layers was 1, the number of filters was set to 128, and the window sizes were 2, 3, and 4. In addition, the LSTM model was used in the experiment, and the output dimension of the hidden units was set as 128. For the combination of global semantic features and syntactic dependency features, the dropout rate was 0.5. The batch size was set to 8. Finally, we used the DL framework Pytorch [57] to implement our model and carry out the experimental process.

Ethics Approval

The data set and methods used in this work are publicly available and do not involve any ethical or moral issues.

Results

Data Set and Evaluation Metrics

We conducted a series of experiments on the BB-rel and SeeDev-binary task data sets to evaluate our proposed approach.

The BB-rel task in BioNLP-OST 2019 is quite different from the previous versions, which integrate the new entity type of *phenotype* and relation category of *Exhibits*. Therefore, this task involves 4 entity types, *microorganism*, *habitat*, *geography*, and *phenotype*, and 2 relation categories between entity pairs, *Lives_In* and *Exhibits*. In practice, the nonrelation between entity pairs is also regarded as a prediction category, so this task is treated as a multi-classification relation extraction task. In addition to intrasentence relations, the BB-rel task also considers intersentence relations, which remains a significant challenge. The proportion of intersentence relationships in the corpus was 17.5%. In our method, we consider only the intrasentence relationship. We adopted the method described in the data preprocessing section to segment the text into sentences, construct negative instances, and remove instances that do not comply with the constraint of regulation. In this manner, we constructed 1996 training instances, including 943 related instances; 1040 validation instances, including 517 related instances; and 1414 test instances. The detailed distribution of the BB-rel task data set after the preprocessing

procedure is summarized in Table 1. Owing to different data revision and processing methods, the number of instances may be inconsistent with other studies.

We used the predictions of the test set to evaluate our methods on the web-based evaluation service [58]. Its evaluation metrics are similar to those of previous versions, including precision, recall, F_1 score, and the results of the intrasentence and intersentence relations of various relation categories [35].

The SeeDev-binary task corpus is a set of 87 paragraphs from 20 full articles on the seed development of *Arabidopsis thaliana*, with 17 entity types and 22 relation categories manually annotated by domain experts. There are 3575 annotated relations, including 1628 relations for the training sets, 819 relations for the validation sets, and 1128 relations for the test sets. We used the same method to preprocess the data set and eliminate intersentence relations. Then, 18,997 training instances were constructed, including 1508 related instances; 8955 validation instances were constructed, including 746 related instances; and 12,737 test instances were constructed, and the detailed distribution is shown in Table 2. It can be seen that there is an extreme imbalance where the number of nonrelation samples far exceeds the positive samples, which is more challenging and will negatively affect the performance of the model [47]. Therefore, to alleviate this problem, through a series of pre-experiments, we finally decided to randomly delete 90% (15,740/17,489) of the negative samples in the training stage, but the validation and test sets were not reduced.

The SeeDev-binary is also applicable to the web-based evaluation services. Compared with SeeDev-binary 2016, task organizers have added new evaluation metrics to emphasize biomedical contributions. The evaluation metrics are global results for all relations, the results of intrasentence relations, and type clusters, each of which has a precision, recall, and F_1 score.

Table 1. Detailed statistics of the relation extraction of Bacteria-Biotope task data set. The statistics of the test set is none because the organizer has not released the annotated relation on the test set.

Category	Training set	Validation set	Test set
Total	1996	1040	1414
Lives_in	659	377	None
Exhibits	284	140	None
Lives_in and Exhibits	943	517	None
Nonrelation	1053	523	None

Table 2. Detailed statistics of the binary relation extraction of plant seed development task data set. The number of relationships in the test set is none because the number of relationships cannot be determined after preprocessing.

Category	Training set	Validation set	Test set
Total	18,997	8955	12,737
All relation	1508	746	None
Nonrelation	17,489	8209	None

Experiment Results

In the BB-rel task, we used the proposed DL model based on the multi-head attention mechanism and syntactic dependency feature to detect biomedical relations. Our proposed method finally obtained an F_1 score of 65.56% on the test set; the details are shown in Table 3. Our method has an F_1 scores of 62.36% and 73.62% for the relation category of *Lives_In* and *Exhibits*, respectively, and performs better in the relation category *Exhibits*. Moreover, it can be noted that the F_1 scores in the identification of intrasentence relations of Lives_In and Exhibits are 69.00% and 77.67%, which are higher than the comprehensive F_1 score. This is because our preprocessing method only deals with intrasentence relations; therefore, it performs better in the identification of intrasentence relations.

Table 4 lists the comparison between our method and other previous systems in BB-rel task. The first 3 lines in the table are the official top 3 systems (10 participated), among which Yuhang_Wu used a multilayer perceptron [35], AliAI [39] used a multitask architecture similar to BERT, and whunlp [37] achieves state-of-the-art performance by using dependency graph and attention graph convolution neural network. The fourth line is the baseline provided by the task organizer, which uses a co-occurrence method. Owing to the huge difference between the model architecture of these systems, only the final F_1 score is used for comparison. The F_1 score of our method is 5.07% higher than the third-placed Yuhang_Wu and 0.60% superior to the second-placed AliAI, who achieved the result of 64.96%. It is worth noting that our model achieved the best precision of 69.50%, which is superior to all existing systems in BB-rel task. This result reveals that our method tends to predict fewer positive classes, that is, it performs better on false positives than other models. In conclusion, this comparison indicates that our proposed model is effective and achieved excellent performance in BB-rel task.

In the SeeDev-binary task, our proposed method achieved an F_1 score of 38.04% for all relations in the test set. The detailed results for the specific relation categories are shown in Table 5. As shown in the table, 7 types of relation categories were not detected, such as *Is_Involved_In_Process* and *Occurs_During*.

Through the statistical analysis of the data set, it was found that there were few positive instances of these relation categories in the training set, which was obviously responsible for the uneven classification.

Table 6 lists the results of comparison between our method and other systems for the SeeDev-binary task. The first 2 systems are the top 2 of the official ranks in BioNLP-OST 2019. Among them, the first-placed MIC-CIS [42] used linguistic feature and SVM classifier to achieve an F_1 score of 37.38%, whereas YNU-junyi [14], the second-ranking system, obtained an F_1 score of 34.18% using a DL model combined with distributed representation, CNN and LSTM model. The results show that our method achieves the state-of-the-art performance in both category of all relation and intrasentence relation, with F_1 scores of 38.04% and 38.68%, respectively. In the all-relation category, the F_1 score of our system outperformed the first-ranking system by 0.66% and the second-ranking system by 3.86%. Meanwhile, the result is similar to BB-rel task; our system performed excellently in precision. In All relation and intrasentence relation, the precision surpassed the first-ranking system by 7.30% and 5.30%, respectively. This once again proves that our model has a lower false-positive rate than other models. Therefore, we can conclude that our model can take advantage of both the multi-head attention mechanism and syntactic dependency feature to achieve excellent performance in biomedical relation extraction tasks.

The results by cluster are also important evaluation metrics in the SeeDev-binary task, and the comparison of F_1 scores is shown in Table 7. It can be seen from the table that our model achieves optimal results in 3 cluster categories: *function*, *regulation*, and *genic regulation*, and it performs poorly in 2 cluster categories: *composition membership* and *interaction*, but the overall performance of our proposed model is generally satisfactory.

Table 3. Detailed results of our method on the test set of relation extraction of Bacteria-Biotope task.

Category	Precision	Recall	F_1 score
Lives_In and Exhibits	69.50	62.05	*65. 56* [a]
Lives_In	69.38	56.64	62.36
Lives_In (intrasentence)	69.75	68.27	69.00
Exhibits	69.77	77.92	73.62
Exhibits (intrasentence)	70.18	86.96	77.67

[a]The final F_1 score is shown in italics.

Table 4. Comparison of results between our method and other systems for the relation extraction of Bacteria-Biotope task.

Models	Precision	Recall	F_1 score
whunlp [37]	62.94	*70.22* [a]	*66.38*
AliAI [39]	68.20	62.01	64.96
Yuhang_Wu [35]	55.10	67.03	60.49
Baseline [35]	52.54	80.13	63.47
Our model	*69.50*	62.05	65.56

[a]The maximum results are shown in italics.

Table 5. Detailed results of our method on the test set of the binary relation extraction of plant seed development task.

Binary relation type	Precision	Recall	F_1 score
Exists_In_Genotype	40.59	32.28	35.96
Occurs_In_Genotype	0	0	0
Exists_At_Stage	50.00	10.00	16.67
Occurs_During	0	0	0
Is_Localized_In	38.16	46.77	42.03
Is_Involved_In_Process	0	0	0
Transcribes_Or_Translates_To	0	0	0
Is_Functionally_Equivalent_To	60.94	55.71	58.21
Regulates_Accumulation	66.67	25.00	36.36
Regulates_Development_Phase	22.86	41.56	29.49
Regulates_Expression	24.65	50.72	33.18
Regulates_Molecule_Activity	0	0	0
Regulates_Process	40.04	64.71	49.47
Regulates_Tissue_Development	0	0	0
Composes_Primary_Structure	60.00	37.50	46.15
Composes_Protein_Complex	50.00	66.67	57.14
Is_Protein_Domain_Of	26.09	19.35	22.22
Is_Member_Of_Family	27.78	52.33	36.29
Has_Sequence_Identical_To	100.00	47.73	64.62
Interacts_With	80.00	14.81	25.00
Binds_To	30.77	12.50	17.78
Is_Linked_To	0	0	0
All relations	34.75	42.02	*38.04* [a]

[a]The final F_1 score is shown in italics.

Table 6. Comparison of results between our method and other systems for the binary relation extraction of plant seed development task.

Models	All relation			Intrasentence relation		
	Precision	Recall	F_1 score	Precision	Recall	F_1 score
MIC-CIS [42]	27.45	*51.15* [a]	37.38	29.45	*53.08*	37.88
YNU-junyi [14]	27.25	45.83	34.18	27.25	47.56	34.65
Our method	*34.75*	42.02	*38.04*	*34.75*	43.61	*38.68*

[a]The maximum results are shown in italics.

Table 7. Comparison of F_1 scores by cluster between our method and other systems for the binary relation extraction of plant seed development task.

Models	All	Comparison	Function	Regulation	Genic regulation	Composition membership	Interaction
MIC-CIS [42]	37.38	47.92	17.39	34.78	33.84	*40.25* [a]	*34.24*
YNU-junyi [14]	34.18	*50.45*	25.00	34.21	23.00	34.68	21.87
Our method	*38.04*	49.68	*25.53*	*40.78*	*34.04*	32.72	22.02

[a]The maximum results are shown in italics.

Discussion

Overview

In this section, we construct ablation experiments to analyze the effectiveness of multi-head attention mechanism and syntactic dependency feature. To avoid the instability of a single model, the mean F_1 score on the test set was used to measure model performance. Subsequently, we conducted an error analysis and manually analyzed the correct and incorrect predictions.

Effectiveness of Multi-Head Attention Mechanism

We first analyzed the effectiveness of the multi-head attention mechanism in the global semantic feature extraction of a full sentence compared with the traditional CNN, BiLSTM, and additive attention models [26]. All models use the distributed features and syntactic dependency features that we use, such as domain-specific word embedding. Owing to the application of PE in the multi-head attention mechanism, we integrate PE into all models for a fair comparison. Table 8 shows a comparison of the mean F_1 scores using various models to encode global semantic features.

From the table, the first 2 lines are the results of extracting the feature representation of sentences using the CNN or BiLSTM model alone, among which the result of the BiLSTM model was slightly better. A possible explanation is that the length of sentences in instances is generally large, and the CNN model can only process window information and rely on a pooling operation to summarize the overall structure of the sentences. However, the BiLSTM model is more suitable for sequence modeling and encoding longer sequence information using a bidirectional memory network. They were then combined with an additive attention model. Compared with CNN and LSTM models alone, the application of the attention model improved F_1 scores by 1.82% and 1.22% on BB-rel and 1.31% and 1.11% on SeeDev-binary, respectively. In addition, the performance of CNN with attention exceeds that of BiLSTM with attention on the BB-rel task, possibly because the attention mechanism fills the shortcoming that CNN cannot capture the long-range dependence of sentences. Hence, these results suggest that the attention mechanism can effectively improve the performance of the model by focusing on the key information of the token sequence and learning the overall structure of a sentence.

Finally, the multi-head attention mechanism is introduced into our model without any CNN or recurrent neural network structure, and the optimal result is achieved. The mean F_1 score was 63.13% and 36.37% for the 2 tasks, which are 1.11% and 1.24% higher than that of the BiLSTM-attention model and 0.96% and 1.45% higher than that of the CNN-attention model, respectively. The results show that the multi-head attention mechanism significantly outperforms the additive attention model in biomedical relation extraction. To some extent, additive attention can be understood as a single-head attention model that can only learn the global semantic features in one representation space. However, the advantage of the multi-head attention mechanism is that it captures the global semantic information in different representation subspaces and integrates the contextual information of relevant words into the current word from multiple channels. The experimental results demonstrate that the multi-head attention mechanism can extract more comprehensive feature representations and effectively improve the performance of the relation extraction model.

Table 8. The comparison of mean F_1 score of using different models to extract global semantic features in the relation extraction of Bacteria-Biotope task (BB-rel) and the binary relation extraction of plant seed development task (SeeDev-binary).

Global semantic features	BB-rel			SeeDev-binary		
	Minimum[a]	Maximum[b]	Mean (SD)	Minimum[a]	Maximum[b]	Mean (SD)
CNN[c]	57.26	63.26	60.35 (2.11)	31.67	35.85	33.61 (1.33)
BiLSTM[d]	57.89	63.80	60.80 (1.88)	32.39	36.28	34.02 (1.53)
CNN-attention	59.69	65.01	62.17 (1.69)	32.89	37.52	34.92 (1.47)
BiLSTM-attention	59.80	64.38	62.02 (1.45)	33.61	37.30	35.13 (1.18)
Multi-head attention	*60.68* [e]	*65.56*	*63.13 (1.55)*	*34.47*	*38.04*	*36.37 (1.13)*

[a]The lowest F_1-scores of 10 different random initializations.

[b]The highest F_1-scores of 10 different random initializations.

[c]CNN: convolutional neural network.

[d]BilSTM: bidirectional long short-term memory network.

[e]The maximum results are shown in italics.

Effectiveness of Syntactic Dependency Feature

Furthermore, we analyzed the effectiveness of the syntactic dependency feature in our model. The length of the shortest dependency paths, based on syntactic analysis, is mostly 2 to 5, which belongs to a super-short sequence. Therefore, we only tried to use the CNN and BiLSTM models for feature extraction, and the results are shown in Table 9. The first line shows the results that the model does not use syntactic dependency features, and the average F_1 scores were 60.85% and 34.60% for BB-rel and SeeDev-binary tasks, respectively. When the LSTM model was used to extract syntactic dependency features,

the mean F_1 scores of the model were 62.88% and 36.06%. When we used the CNN model, the performance of the model reached optimal F_1 scores, which improved to 63.13% and 36.37% on BB-rel and SeeDev-binary tasks, respectively. The results also show that the CNN model is superior to LSTM in terms of feature extraction for super-short sequences. By comparison, it can be demonstrated that the integration of syntactic dependency features can enable the model to learn syntactic information between entity pairs through a dependency graph, which can effectively improve the performance of the model.

Table 9. The comparison of mean F_1 scores of using different models to extract syntactic dependency features in the relation extraction of Bacteria-Biotope task (BB-rel) and the binary relation extraction of plant seed development task (SeeDev-binary).

Syntactic dependency feature	BB-rel			SeeDev-binary		
	Minimum[a]	Maximum[b]	Mean (SD)	Minimum[a]	Maximum[b]	Mean (SD)
No-use	58.51	63.70	60.85 (1.65)	32.89	36.53	34.60 (1.16)
LSTM[c]	59.93	65.16	62.88 (1.66)	*34.55* [d]	37.90	36.06 (1.07)
CNN[e]	*60.68*	*65.56*	*63.13 (1.55)*	34.47	*38.04*	*36.37 (1.13)*

[a]The lowest F_1-scores of 10 different random initializations.

[b]The highest F_1-scores of 10 different random initializations.

[c]LSTM: long short-term memory network.

[d]The maximum results are shown in italics.

[e]CNN: convolutional neural network.

Error Analysis

To verify the advantages and weaknesses of our proposed model, we compared the experimental results with those of other existing models. We find that our system performs better in terms of the precision of the 2 relation extraction tasks, far surpassing other models, which means that our approach has a lower false-positive rate than the other models. One possible explanation is that our model structure introduces the shortest dependent paths compared with other systems, which can more definitely identify the biomedical relationship between entity pairs.

The 2 relationship extraction tasks are constrained under regulations; therefore, it is necessary to investigate whether there is a situation in which the predicted relationship does not conform to the rules. For example, in the sentence "An evaluation of selective broths based on the bi-selenite ion and on hypertonic strontium chloride in *Salmonellae* detection in egg products," the entity "Salmonellae" is of *microorganism* type, and the entity "egg products" is of *habitat* type. There may be a *Lives_In* relationship between them, but if it is predicted as *an Exhibits* relationship, it must be wrong. Through an analysis of the prediction results on the validation set, it was found that this situation rarely occurs. Therefore, our research should focus on whether a biomedical relationship exists between entity pairs.

In addition, we manually analyzed the correct and false predictions from the validation set compared with existing DL models (structures similar to YNU-junyi [14]). We found that our proposed model generally performed better on long sentences. A complicated sentence structure and long distance between 2 entities are more likely to lead to relationship classification errors. For example, in the sentence "The prevalence of H. pylori infection in dyspeptic patients in Yemen is very high, the eradication rate with standard triple therapy was unsatisfactory probably because of widespread bacterial resistance due to unrestricted antibiotic use," "H. pylori" is a *microorganism* entity, "widespread bacterial resistance due to unrestricted antibiotic use" is *a phenotypic* entity, and there is an *Exhibits* relationship between them. The DL model, similar to YNU-junyi, predicted it as a nonrelationship category, but our model can better detect it, probably because our proposed model can capture the long-term dependency between words in a long sentence.

Conclusions

This paper focuses on the 2 relation extraction tasks in BioNLP-OST 2019: BB-rel task and SeeDev-binary task, which aim to promote the development of fine-grained IE from biomedical texts. For these tasks, we propose a DL model based on the combination of a series of distributed features to detect relations, introduce a multi-head attention mechanism to extract global semantic features, and use syntactic-dependent features to enrich the feature representation. Our proposed method obtained F_1 scores of 65.56% and 38.04% on the test sets of the 2 tasks and achieved state-of-the-art results in the SeeDev-binary task. Through ablation experiments, the effectiveness of multi-head attention and syntactic dependency features was demonstrated. The multi-head attention mechanism allows the model to learn relevant semantic information in different representation subspaces at different positions and integrates the contextual information of relevant words in the sentence into the current word representation, which greatly improves the performance of the biomedical relation extraction model.

Despite the excellent performance of our model on BB-rel and SeeDev-binary tasks, there are still many challenges. In particular, the intersentence relation is not considered in our method, which remains a difficult problem in biomedical relation extraction tasks. This situation is because of the complexity of the reasoning relationship and the extreme imbalance between

the positive and negative examples. In contrast, the use of a DL model to extract high-quality features from small training data sets is a problem that needs to be solved. In future work, we will consider using a semisupervised learning method or transformer model, such as BERT, to better solve the topic of biomedical relation extraction.

Acknowledgments

This study was supported by the Youth Science and Technology Talent Growth Project of the general university in Guizhou Province (黔教合KY字 [2022] 281号), the Zunyi Science and Technology Cooperation Fund (遵市科合HZ字 [2020] 81号), and the Guizhou Science and Technology Cooperation Platform Talent Fund (黔科合平台人才 [2018] 5772-088, 黔科合平台人才 [2019]-020).

Conflicts of Interest

None declared.

References

1. Mooney RJ, Bunescu R. Mining knowledge from text using information extraction. SIGKDD Explor Newsl 2005 Jun 01;7(1):3-10. [doi: 10.1145/1089815.1089817]
2. Krallinger M, Erhardt RA, Valencia A. Text-mining approaches in molecular biology and biomedicine. Drug Discov Today 2005 Mar 15;10(6):439-445. [doi: 10.1016/S1359-6446(05)03376-3] [Medline: 15808823]
3. Zweigenbaum P, Demner-Fushman D, Yu H, Cohen KB. Frontiers of biomedical text mining: current progress. Brief Bioinform 2007 Sep;8(5):358-375 [FREE Full text] [doi: 10.1093/bib/bbm045] [Medline: 17977867]
4. Blaschke C, Andrade MA, Ouzounis C, Valencia A. Automatic extraction of biological information from scientific text: protein-protein interactions. Proc Int Conf Intell Syst Mol Biol 1999:60-67. [Medline: 10786287]
5. Segura-Bedmar I, Martínez P, de Pablo-Sánchez C. Extracting drug-drug interactions from biomedical texts. BMC Bioinformatics 2010 Oct 06;11(S5):P9 [FREE Full text] [doi: 10.1186/1471-2105-11-s5-p9]
6. Nédellec C, Bossy R, Kim JD. Proceedings of the 4th BioNLP Shared Task Workshop. 2016 Presented at: BioNLP '16; August 13, 2016; Berlin, Germany. [doi: 10.18653/v1/w16-30]
7. BioNLP Open Shared Tasks 2019. URL: https://2019.bionlp-ost.org/home [accessed 2022-09-01]
8. Kambhatla N. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In: Proceedings of the ACL 2004 on Interactive poster and demonstration sessions. 2004 Presented at: ACLdemo '04; July 21-26, 2004; Barcelona, Spain p. 22-es. [doi: 10.3115/1219044.1219066]
9. Nguyen TH, Grishman R. Employing word representations and regularization for domain adaptation of relation extraction. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2014 Presented at: ACL '14; June 22-27, 2014; Baltimore, MD, USA p. 68-74. [doi: 10.3115/v1/p14-2012]
10. Nguyen TV, Moschitti A, Riccardi G. Convolution kernels on constituent, dependency and sequential structures for relation extraction. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. 2009 Aug Presented at: EMNLP '09; August 6-7, 2009; Singapore, Singapore p. 1378-1387. [doi: 10.3115/1699648.1699684]
11. Sun L, Han X. A feature-enriched tree kernel for relation extraction. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2014 Presented at: ACL '14; June 22-27, 2014; Baltimore, MD, USA p. 61-67. [doi: 10.3115/v1/p14-2011]
12. Björne J, Salakoski T. TEES 2.1: Automated annotation scheme learning in the BioNLP 2013 Shared Task. In: Proceedings of the BioNLP Shared Task 2013 Workshop. 2013 Presented at: BioNLP '13; August 9, 2013; Sofia, Bulgaria p. 16-25 URL: https://aclanthology.org/W13-2003.pdf
13. Li H, Zhang J, Wang J, Lin H, Yang Z. DUTIR in BioNLP-ST 2016: utilizing convolutional network and distributed representation to extract complicate relations. In: Proceedings of the 4th BioNLP shared task workshop. 2016 Presented at: BioNLP '16; August 13, 2016; Berlin, Germany p. 93-100 URL: https://aclanthology.org/W16-3012.pdf
14. Li J, Zhou X, Wu Y, Wang B. YNU-junyi in BioNLP-OST 2019: Using CNN-LSTM Model with Embeddings for SeeDev Binary Event Extraction. In: Proceedings of The 5th Workshop on BioNLP Open Shared Tasks. 2019 Presented at: BioNLP '19; November 4, 2019; Hong Kong, China p. 110-114. [doi: 10.18653/v1/D19-5717]
15. LeCun Y, Bengio Y. Convolutional networks for images, speech, and time series. In: Arbib MA, editor. The Handbook of Brain Theory and Neural Networks. Cambridge, MA, USA: MIT Press; Oct 1998:255-258.
16. Liu S, Tang B, Chen Q, Wang X. Drug-drug interaction extraction via convolutional neural networks. Comput Math Methods Med 2016;2016:6918381 [FREE Full text] [doi: 10.1155/2016/6918381] [Medline: 26941831]
17. Zeng D, Liu K, Lai S, Zhou G, Zhao J. Relation classification via convolutional deep neural network. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. 2014 Presented at: COLING '14; August 23-29, 2014; Dublin, Ireland p. 2335-2344 URL: https://aclanthology.org/C14-1220.pdf

18. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997 Nov 15;9(8):1735-1780. [doi: 10.1162/neco.1997.9.8.1735] [Medline: 9377276]
19. Zhang Y, Zheng W, Lin H, Wang J, Yang Z, Dumontier M. Drug-drug interaction extraction via hierarchical RNNs on sequence and shortest dependency paths. Bioinformatics 2018 Mar 01;34(5):828-835 [FREE Full text] [doi: 10.1093/bioinformatics/btx659] [Medline: 29077847]
20. Sahu SK, Anand A. Drug-drug interaction extraction from biomedical texts using long short-term memory network. J Biomed Inform 2018 Oct;86:15-24 [FREE Full text] [doi: 10.1016/j.jbi.2018.08.005] [Medline: 30142385]
21. Vu NT, Adel H, Gupta P, Schütze H. Combining recurrent and convolutional neural networks for relation classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016 Jun Presented at: NAACL '16; June 12-17, 2016; San Diego, California p. 534-539 URL: https://aclanthology.org/N16-1065/
22. Zheng S, Hao Y, Lu D, Bao H, Xu J, Hao H, et al. Joint entity and relation extraction based on a hybrid neural network. Neurocomputing 2017 Sep;257:59-66. [doi: 10.1016/j.neucom.2016.12.075]
23. Pyysalo S, Ginter F, Moen H, Salakoski T, Ananiadou S. Distributional semantics resources for biomedical text processing. In: Proceedings of the 5th International Symposium on Languages in Biology and Medicine. 2013 Presented at: LBM '13; December 12-13, 2013; Tokyo, Japan p. 39-44 URL: https://bio.nlplab.org/pdf/pyysalo13literature.pdf
24. Pasupa K, Seneewong Na Ayutthaya T. Thai sentiment analysis with deep learning techniques: a comparative study based on word embedding, POS-tag, and sentic features. Sustain Cities Soc 2019 Oct;50:101615. [doi: 10.1016/j.scs.2019.101615]
25. Cormode G. Sequence distance embeddings. Department of Computer Science, The University of Warwick. 2003 Jan. URL: https://www.dcs.warwick.ac.uk/report/pdfs/cs-rr-393.pdf [accessed 2022-09-01]
26. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. In: Proceedings of the 3rd International Conference on Learning Representations. 2015 Presented at: ICLR '15; May 7-9, 2015; San Diego, CA, USA URL: https://arxiv.org/abs/1409.0473
27. Luong MT, Pham H, Manning CD. Effective approaches to attention-based neural machine translation. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015 Sep Presented at: EMNLP '15; September 17-21, 2015; Lisbon, Portugal p. 1412-1421 URL: https://aclanthology.org/D15-1166/
28. Yu AW, Dohan D, Luong MT, Zhao R, Chen K, Norouzi M, et al. QANet: combining local convolution with global self-attention for reading comprehension. In: Proceedings of the 6th International Conference on Learning Representations. 2018 Presented at: ICLR '18; April 30-May 3, 2018; Vancouver, Canada URL: https://openreview.net/forum?id=B14TlG-RW
29. Wang Y, Huang M, Zhu X, Zhao L. Attention-based LSTM for aspect-level sentiment classification. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 2016 Presented at: EMNLP '16; November 1-5, 2016; Austin, TX, USA p. 606-615 URL: https://aclanthology.org/D16-1058.pdf
30. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Joens L, Gomez AN, et al. Attention is all you need. In: Proceedings of the 2017 Advances in Neural Information Processing Systems. 2017 Presented at: NeurIPS '17; December 4-9, 2017; Long Beach, CA, USA p. 5998-6008.
31. Mehryary F, Björne J, Pyysalo S, Salakoski T, Ginter F. Deep learning with minimal training data: TurkuNLP entry in the BioNLP shared task 2016. In: Proceedings of the 4th BioNLP Shared Task Workshop. 2016 Presented at: BioNLP '16; August 13, 2016; Berlin, Germany p. 73-81 URL: https://aclanthology.org/W16-3009.pdf [doi: 10.18653/v1/W16-3009]
32. Hua L, Quan C. A shortest dependency path based convolutional neural network for protein-protein relation extraction. Biomed Res Int 2016;2016:8479587 [FREE Full text] [doi: 10.1155/2016/8479587] [Medline: 27493967]
33. Jettakul A, Wichadakul D, Vateekul P. Relation extraction between bacteria and biotopes from biomedical texts with attention mechanisms and domain-specific contextual representations. BMC Bioinformatics 2019 Dec 03;20(1):627 [FREE Full text] [doi: 10.1186/s12859-019-3217-3] [Medline: 31795930]
34. Deléger L, Bossy R, Chaix E, Ba M, Ferré A, Bessières P, et al. Overview of the bacteria biotope task at BioNLP shared task 2016. In: Proceedings of the 4th BioNLP Shared Task Workshop. 2016 Presented at: BioNLP '16; August 13, 2016; Berlin, Germany p. 12-22. [doi: 10.18653/v1/w16-3002]
35. Bossy R, Deléger L, Chaix E, Ba M, Nédellec C. Bacteria biotope at BioNLP open shared tasks 2019. In: Proceedings of The 5th Workshop on BioNLP Open Shared Tasks. 2019 Presented at: BioNLP '19; November 4, 2019; Hong Kong, China p. 121-131. [doi: 10.18653/v1/D19-5719]
36. Li L, Wan J, Zheng J, Wang J. Biomedical event extraction based on GRU integrating attention mechanism. BMC Bioinformatics 2018 Aug 13;19(Suppl 9):285 [FREE Full text] [doi: 10.1186/s12859-018-2275-2] [Medline: 30367569]
37. Xiong W, Li F, Cheng M, Yu H, Ji D. Bacteria biotope relation extraction via lexical chains and dependency graphs. In: Proceedings of The 5th Workshop on BioNLP Open Shared Tasks. 2019 Presented at: BioNLP '19; November 4, 2019; Hong Kong, China p. 158-167. [doi: 10.18653/v1/D19-5723]
38. Schuster M, Paliwal KK. Bidirectional recurrent neural networks. IEEE Trans Signal Process 1997 Nov;45(11):2673-2681. [doi: 10.1109/78.650093]
39. Zhang Q, Liu C, Chi Y, Xie X, Hua X. A multi-task learning framework for extracting bacteria biotope information. In: Proceedings of The 5th Workshop on BioNLP Open Shared Tasks. 2019 Presented at: BioNLP '19; November 4, 2019; Hong Kong, China p. 105-109 URL: https://aclanthology.org/D19-5716/ [doi: 10.18653/v1/D19-5716]

40. Chaix E, Dubreucq B, Fatihi A, Valsamou D, Bossy R, Ba M, et al. Overview of the regulatory network of plant seed development (SeeDev) task at the BioNLP shared task 2016. In: Proceedings of the 4th BioNLP Shared Task Workshop. 2016 Presented at: BioNLP '16; August 13, 2016; Berlin, Germany p. 1-11. [doi: 10.18653/v1/W16-3001]
41. Panyam NC, Khirbat G, Verspoor K, Cohn T, Ramamohanarao K. SeeDev binary event extraction using SVMs and a rich feature set. In: Proceedings of the 4th BioNLP Shared Task Workshop. 2016 Presented at: BioNLP '16; August 13, 2016; Berlin, Germany p. 82-87. [doi: 10.18653/v1/W16-3010]
42. Gupta P, Yaseen U, Schütze H. Linguistically informed relation extraction and neural architectures for nested named entity recognition in BioNLP-OST 2019. In: Proceedings of The 5th Workshop on BioNLP Open Shared Tasks. 2019 Presented at: BioNLP '19; November 4, 2019; Hong Kong, China p. 132-142. [doi: 10.18653/v1/D19-5720]
43. Kim JD, Ohta T, Tateisi Y, Tsujii J. GENIA corpus--semantically annotated corpus for bio-textmining. Bioinformatics 2003;19 Suppl 1:i180-i182. [doi: 10.1093/bioinformatics/btg1023] [Medline: 12855455]
44. Charniak E, Johnson M. Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. 2005 Presented at: ACL '05; June 25-30, 2005; Ann Arbor, MI, USA p. 173-180 URL: https://aclanthology.org/P05-1022.pdf [doi: 10.3115/1219840.1219862]
45. McClosky D. Any Domain Parsing: Automatic Domain Adaptation for Natural Language Parsing. Providence, RI, USA: Brown University; 2010.
46. de Marneffe MC, MacCartney B, Manning CD. Generating typed dependency parses from phrase structure parses. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation. 2006 Presented at: LRE '06; May 22-28, 2006; Genoa, Italy p. 449-454 URL: http://www.lrec-conf.org/proceedings/lrec2006/pdf/440_pdf.pdf
47. Ye W, Li B, Xie R, Sheng Z, Chen L, Zhang S. Exploiting entity BIO tag embeddings and multi-task learning for relation extraction with imbalanced data. In: Proceedings of the 57th Conference of the Association for Computational Linguistics. 2019 Presented at: ACL '19; July 28- August 2, 2019; Florence, Italy p. 1351-1360 URL: https://aclanthology.org/P19-1130/
48. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2. 2013 Presented at: NIPS '13; December 5-10, 2013; Lake Tahoe, NV, USA p. 3111-3119.
49. Zheng W, Lin H, Luo L, Zhao Z, Li Z, Zhang Y, et al. An attention-based effective neural model for drug-drug interactions extraction. BMC Bioinformatics 2017 Oct 10;18(1):445 [FREE Full text] [doi: 10.1186/s12859-017-1855-x] [Medline: 29017459]
50. Liu S, Shen F, Komandur Elayavilli R, Wang Y, Rastegar-Mojarad M, Chaudhary V, et al. Extracting chemical-protein relations using attention-based neural networks. Database (Oxford) 2018 Jan 01;2018:bay102 [FREE Full text] [doi: 10.1093/database/bay102] [Medline: 30295724]
51. Bunescu R, Mooney R. A shortest path dependency kernel for relation extraction. In: Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing. 2005 Presented at: EMNLP '05; October 6-8, 2005; Vancouver, Canada p. 724-731 URL: https://aclanthology.org/H05-1091.pdf
52. Chowdhury FM, Lavelli A, Moschitti A. A study on dependency tree kernels for automatic extraction of protein-protein interaction. In: Proceedings of BioNLP 2011 Workshop. 2011 Presented at: BioNLP '11; June 23-24, 2011; Portland, OR, USA p. 124-133 URL: https://aclanthology.org/W11-0216.pdf
53. Xu Y, Mou L, Li G, Chen Y, Peng H, Jin Z. Classifying relations via long short term memory networks along shortest dependency paths. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015 Presented at: EMNLP '15; September 17-21, 2015; Lisbon, Portugal p. 1785-1794 URL: https://aclanthology.org/D15-1206.pdf
54. Kim Y. Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. 2014 Oct Presented at: EMNLP '16; October 25-29, 2014; Doha, Qatar p. 1746-1751 URL: https://aclanthology.org/D14-1181/
55. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 2014;15(56):1929-1958.
56. Kingma DP, Ba J. Adam: a method for stochastic optimization. In: Proceedings of the 3rd International Conference on Learning Representations. 2014 Dec 22 Presented at: ICLR '15; May 7-9, 2015; San Diego, CA, USA URL: https://arxiv.org/abs/1412.6980
57. Paszke A, Gross S, Massa F, Lerer A, Bradbury H, Chanan G, et al. PyTorch: an imperative style, high-performance deep learning library. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. 2019 Presented at: NeurIPS '19; December 8-14, 2019; Vancouver, Canada p. 8024-8035 URL: https://proceedings.neurips.cc/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf
58. BioNLP-OST 2019 Evaluation Service. Institut National de la Recherche Agronomique. 2019. URL: http://bibliome.jouy.inra.fr/demo/BioNLP-OST-2019-Evaluation/index.html [accessed 2022-06-01]

Abbreviations

BB-rel: relation extraction of Bacteria-Biotope task

BERT: Bidirectional Encoder Representations from Transformers
BiLSTM: bidirectional long short-term memory
BioNLP-OST: Biomedical Natural Language Processing Workshop-Open Shared Task
CNN: convolutional neural network
DL: deep learning
IE: information extraction
LSTM: long short-term memory
PE: position embedding
POS: part of speech
SeeDev-binary: binary relation extraction of plant seed development task
SVM: support vector machine
TEES: Turku Event Extraction System

Edited by C Lovis, J Hefner; submitted 16.07.22; peer-reviewed by Y Cui, M Wang; comments to author 02.08.22; revised version received 27.08.22; accepted 07.09.22; published 20.10.22.

Please cite as:
Li Y, Hui L, Zou L, Li H, Xu L, Wang X, Chua S
Relation Extraction in Biomedical Texts Based on Multi-Head Attention Model With Syntactic Dependency Feature: Modeling Study
JMIR Med Inform 2022;10(10):e41136
URL: https://medinform.jmir.org/2022/10/e41136
doi:10.2196/41136
PMID:36264604

Original Paper

Identifying Patients With Heart Failure Who Are Susceptible to De Novo Acute Kidney Injury: Machine Learning Approach

Caogen Hong[1,2], MSc; Zhoujian Sun[3], PhD; Yuzhe Hao[2], MSc; Zhanghuiya Dong[2], MSc; Zhaodan Gu[2], MSc; Zhengxing Huang[4], PhD

[1]College of Biomedical Engineering and Instrument Science, Zhejiang University, Hangzhou, China

[2]Jiangsu Automation Research Institute, Lianyungang, China

[3]Research Center for Applied Mathematics and Machine Intelligence, Zhejiang Lab, Hangzhou, China

[4]College of Computer Science and Technology, Zhejiang University, Hangzhou, China

Corresponding Author:
Zhoujian Sun, PhD
Research Center for Applied Mathematics and Machine Intelligence
Zhejiang Lab
Kechuang Ave, Zhongtai Subdistrict, Yuhang District
Hangzhou, 311121
China
Phone: 86 571 56390515
Fax: 86 518 85983716
Email: sunzj@zhejianglab.edu.cn

Abstract

Background: Studies have shown that more than half of patients with heart failure (HF) with acute kidney injury (AKI) have newonset AKI, and renal function evaluation markers such as estimated glomerular filtration rate are usually not repeatedly tested during the hospitalization. As an independent risk factor, delayed AKI recognition has been shown to be associated with the adverse events of patients with HF, such as chronic kidney disease and death.

Objective: The aim of this study is to develop and assess of an unsupervised machine learning model that identifies patients with HF and normal renal function but who are susceptible to de novo AKI.

Methods: We analyzed an electronic health record data set that included 5075 patients admitted for HF with normal renal function, from which 2 phenogroups were categorized using an unsupervised machine learning algorithm called K-means clustering. We then determined whether the inferred phenogroup index had the potential to be an essential risk indicator by conducting survival analysis, AKI prediction, and the hazard ratio test.

Results: The AKI incidence rate in the generated phenogroup 2 was significantly higher than that in phenogroup 1 (group 1: 106/2823, 3.75%; group 2: 259/2252, 11.50%; P<.001). The survival rate of phenogroup 2 was consistently lower than that of phenogroup 1 (P<.005). According to logistic regression, the univariate model using the phenogroup index achieved promising performance in AKI prediction (sensitivity 0.710). The generated phenogroup index was also significant in serving as a risk indicator for AKI (hazard ratio 3.20, 95% CI 2.55-4.01). Consistent results were yielded by applying the proposed model on an external validation data set extracted from Medical Information Mart for Intensive Care (MIMIC) III pertaining to 1006 patients with HF and normal renal function.

Conclusions: According to a machine learning analysis on electronic health record data, patients with HF who had normal renal function were clustered into separate phenogroups associated with different risk levels of de novo AKI. Our investigation suggests that using machine learning can facilitate patient phengrouping and stratification in clinical settings where the identification of high-risk patients has been challenging.

(JMIR Med Inform 2022;10(10):e37484) doi:10.2196/37484

KEYWORDS

heart failure; acute kidney injury; unsupervised machine learning; risk stratification; phenogrouping

Introduction

Acute kidney injury (AKI) is a common disorder in patients with heart failure (HF), with the reported incidence rate varying from 7% to 38% in cardiology departments [1-3]. A recently conducted nationwide survey in China showed that about 85% of AKI incidents that occurred during cardiac hospitalization were ignored or were late to be identified [4,5]. As an independent risk factor, the delayed recognition of AKI has been proven to be associated with worse outcomes of patients with HF (eg, chronic kidney diseaseand mortality) [4,6]. To this end, the prompt identification of patients with HF at high-risk of AKI has great potential to improve clinical outcomes.

Although a few specific clinical markers (eg, estimated glomerular filtration rate [eGFR]) have been adopted to evaluate the renal function of patients with HF such that those at high risk of AKI can be identified, these markers lack the ability to screen de novo AKI patients who had normal renal function at admission [7,8]. Of note, several recently conducted population studies have indicated that more than half of the AKI that occurred in patients with HF were de novo [1-3]. To address this challenge, we attempted to clarify the characteristics of patients with HF who are susceptible to de novo AKI and developed a machine learning model for identification of HF patients with normal renal function but at high risk of de novo AKI.

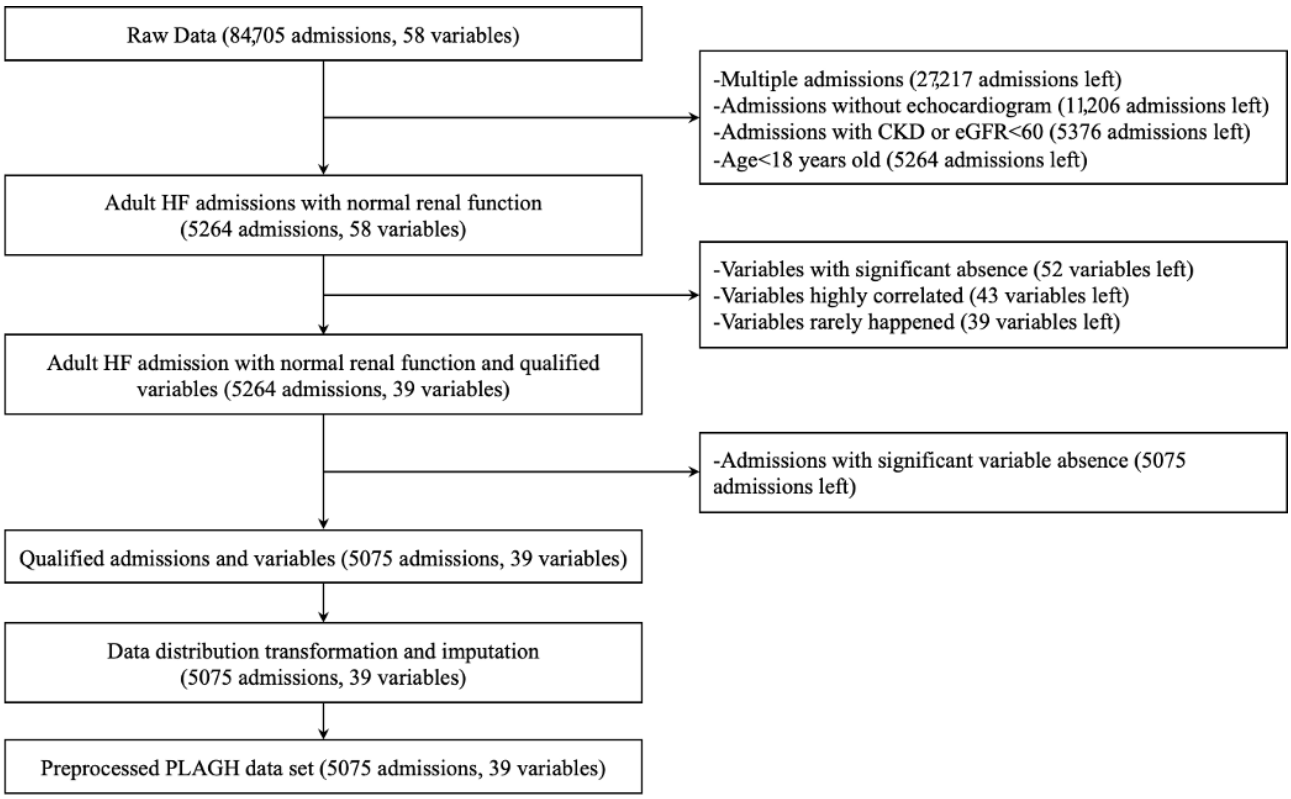
As recently conducted cardiovascular studies have demonstrated that an unsupervised machine learning approach is able to model correlations among variables that contain prognostic information and cluster cohesive patients into 1 homogeneous phenogroup [9-11], we hypothesized that it can also be applied to identify patients with HF at high risk of de novo AKI. Recently, with the rapid development of hospital information systems, a large collection of electronic health records (EHRs) has become available that documents various types of patient information (eg, vital signs, laboratory test results) and treatments (eg, medication, surgery) and thus offers the considerable potential to implement a large-scale real-world analysis at a low expenditure. Therefore, in this study, we aimed to develop an EHR-based unsupervised machine learning analysis to group patients with HF and identify those who are susceptible to de novo AKI.

Methods

Study Population

The proposed retrospective study used a real-world data set obtained from the EHR system of the Chinese PLA General Hospital (PLAGH). The data set documented regular medical information in 84,705 hospitalizations of 29,699 patients who were diagnosed with HF in the PLAGH from 1998 to 2018. Adult patients with HF and normal renal function (eGFR >60 mL/min/1.73m^2 as calculated by the serum creatinine [SCr] version of the Chronic Kidney Disease Epidemiology Collaboration [CKD-EPI] equation [12] and without chronic kidney disease diagnosis) were considered for inclusion. Additionally, patients who did not have echocardiogram records were excluded. For patients with multiple hospitalizations, only the last hospitalization was reserved. The detailed preprocessing procedure is illustrated in Figure 1.

Figure 1. Preprocessing procedure of the PLAGH data set. CKD: chronic kidney disease; eGFR: estimate glomerular filtration rate; HF: heart failure; PLAGH: PLA General Hospital.



Ethics Approval

The study protocol was approved, with a waiver of consent granted on the basis of minimal harm and general impracticability by the health institutional review board of Zhejiang University (No. ZJU-2021-27).

Variable Selection and Machine Learning Model

In this study, 58 variables potentially associated with AKI, including demographics, vital sign measurements, medications, laboratories, operations, and echocardiogram exams, and routinely documented in EHRs at the admission stage of hospitalization were considered as candidates for analysis. To ensure that the most informative variables were selected and the correlation between variables could be diluted, we excluded variables with a missing rate larger than 30% or with a Pearson correlation coefficient >0.6 or that were documented fewer than 100 times in the raw EHR data set. As a result, 39 variables were included in the cohort. All continuous variables were transformed to standard normal distribution for the convenience of the unsupervised machine learning model (Table S1, Multimedia Appendix 1). Thereafter, we adopted multivariate imputation by chained equations [13] to impute the missing data.

We employed a simple yet effective unsupervised machine learning model called K-means clustering to categorize patients into different phenogroups [14]. The silhouette coefficient was applied to determine the optimal number of phenogroups [15]. We also adopted the nonlinear dimensionality reduction technique of t-distributed stochastic neighbor embedding [16] to visualize and evaluate the clustering results in a qualitative manner. The model was repeatedly run 1000 times to guarantee the achieved results stable.

Outcomes of Interest

The primary outcome was the incidence of AKI, which was defined according to the Kidney Disease: Improving Global Outcomes (KDIGO) standard [17], with the occurrence of AKI defined as the increase of SCr to ≥1.5 times the baseline in 7 days or the increase of SCr by ≥26.5 μmol/L within 48 hours. The secondary outcome was in-hospital mortality.

Characterization of Phenogroups

Once patients with HF were categorized into separate phenogroups, we measured the differences of variables in different groups. Continuous variables are reported as median and IQR (interquartile range). Categorical variables are reported as the frequencies and counts. Differences between groups were tested using the 1-way analysis of variance, Kruskal-Wallis test, or the chi-square test where appropriate. A P value of <.01 was considered statistically significant.

Discrimination of Phenogroups

We validated whether the phenogroup index generated by K-means clustering correlated with outcomes of interests by carrying out the following 3 experiments. First, Kaplan-Meier estimators with log-rank tests were conducted to analyze the time-to-event characteristics in different phenogroups. Second, we compared the prediction performance on AKI and in-hospital mortality to check whether the inferred phenogroup index was an effective risk predictor for outcomes of interest. Specifically, we selected the top-ranked 10 variables using a forward stepwise strategy with the Akaike information criterion and then developed 5 logistic regression (LR) models to predict the outcomes of interest. Model 1 used the phenogroup index as the univariate predictor. Model 2 used the top-ranked 10 variables as predictors. Model 3 used the top-ranked 10 variables and the phenogroup index. Model 4 used all 39 variables. Model 5 used all 39 variables and the phenogroup index. All models were trained by 70% of the data from the PLAGH data set and tested with the remaining 30% of data. Third, to evaluate whether the phenogroup index could achieve the competitive discriminative performance compared to the original variables with respect to the primary and secondary outcomes, we applied unadjusted Cox proportional hazard regression to examine hazard ratios (HRs), 95% CIs, and P values for all included original variables as well as the phenogroup index on both the whole PLAGH data set and the following subgroups: age (age <65 vs ≥65 years), sex, type of HF (acute vs chronic), diabetes mellitus, stroke, atrial fibrillation, coronary heart disease, anemia, and left ventricular ejection fraction (<40%, 40%-49%, and ≥50%). To assess continuous variables appropriately, we categorized all continuous variables in validation, and the cutoff points for these continuous variables are presented in online supplementary Table S2, Multimedia Appendix 1.

External Validation

We externally validated our model on a well-known open-source database, Medical Information Mart for Intensive Care (MIMIC)-III [18]. After a requisite preprocessing procedure (online supplementary, Figure S1), we prepared a MIMIC-III data set that contained 1006 patients with HF who had normal renal function. The model trained by the PLAGH data set was directly transferred onto the MIMIC-III data set. In detail, we compared the distance between the data of each patient in the MIMIC-III data set and the centroids of the derived phenogroups from the PLAGH data set and then assigned the patient into a phenogroup with the minimum Euclidean distance. After that, we assessed the survival rate and prediction performance of AKI and in-hospital mortality of the generated phenogroups from the MIMIC-III data set. As patients contained in the PLAGH data set were mainly from general wards in the PLAGH and patients included in the MIMIC-III data set were from intensive care units in the United States, there inevitably were statistical differences between the baseline characteristics of patients in the 2 data sets (Table S3, Multimedia Appendix 1). In this sense, the external validation was able to evaluate the stability of the proposed model in diverse clinical settings.

In this study, statistical and machine learning analysis was based on sklearn, lifelines, scipy package [19-21], and Python. We also report the centroids of the generated phenogroups from the PLAGH data set (Table S4, Multimedia Appendix 1), which may be nontrivial knowledge to assist clinicians in identifying their patients with HF at high risk of de novo AKI.

Results

Phenogroup Results

After preprocessing, 5075 hospitalizations and 39 variables (Table 1) were reserved for the PLAGH data set (median age 61 years, IQR 51-70 years; female 1723/5075, 32.39%; acute HF 1723/5075, 33.95%). Using K-means clustering, we naturally separated patients into 2 basically nonoverlapping phenogroups, where the number of clusters was suggested by the silhouette coefficient test (Figure S1, Multimedia Appendix 1). Similar results were found using t-distributed stochastic neighbor embedding visualization (Figure S2, Multimedia Appendix 1).

Table 1. Included variables for clustering.

Domain	Features
Demographic	Age, sex
Disease	Acute/chronic HF, atrial fibrillation, cardiomyopathy, coronary heart disease, diabetes, stroke, valvular heart disease
Medication	Angiotensin-converting enzyme inhibitor/angiotensin receptor blocker, anticoagulant, antiplatelet, beta blocker, calcium channel blocker, diuretic, positive inotropic drug, vasodilator
Echocardiography	Left ventricular ejection function
Laboratory result	Alanine aminotransferase, aspartate transaminase, estimated glomerular filtration rate, gamma-glutamyl transferase, hemoglobin, high-density lipoprotein cholesterol, low-density lipoprotein cholesterol, N-terminal probrain natriuretic peptide, serum calcium, serum potassium, serum sodium, serum urea, total bilirubin, total serum protein, triglyceride, troponin T
Operation	Angiography percutaneous coronary intervention
Vital sign	BMI, diastolic blood pressure, systolic blood pressure

[a]Only drugs used in the first 48 hours after admission were included to ensure the drug usage could reflect the patient admission status.

Characteristics of Phenogroups

Table 2 illustrates the baseline characteristics of the PLAGH data set and the 2 derived phenogroups. Compared to phenogroup 1, phenogroup 2 had a higher rates of AKI (group 1: 106/2823, 3.75%; group 2: 259/2252, 11.50%; *P*<.001) and in-hospital mortality (phenogroup 1: 21/2823, 0.74%; phenogroup 2: 118/2252, 5.24%; *P*<.001). In addition, patients in phenogroup 2 were generally older than those in phenogroup 1 (58 vs 65 years; *P*<.001).

As can be seen in Table 2, there are more patients diagnosed with acute HF in phenogroup 2 than those in phenogroup 1 (phenogroup 1: 738/2823, 26.14%; phenogroup 2: 985/2252, 43.74%; *P*<.001). Moreover, cardiac function of patients in phenogroup 2 was worse than that in phenogroup 1. Specifically, there were statistical differences between patients in phenogroup 1 and phenogroup 2 in terms of left ventricular ejection fraction (50% vs 41%; *P*<.001), diastolic blood pressure (77 mmHg vs 70 mmHg; *P*<.001), systolic blood pressure (130 mmHg vs 118 mmHg; *P*<.001), N-terminal pro-brain natriuretic peptide (572 pg/mL vs 2680 pg/mL; *P*<.001), hemoglobin (143 g/L vs 129 g/L; *P*<.001), atrial fibrillation (phenogroup 1: 526/2823, 18.63%; phenogroup 2: 595/2252, 26.42%; *P*<.001), diuretic usage (phenogroup 1: 1608/2823, 56.96%; phenogroup 2: 1799/2252, 79.88%; *P*<.001), and positive inotropic drug usage (phenogroup 1: 778/2823, 27.56%; phenogroup 2: 1089/2252, 48.36%; *P*<.001). Furthermore, phenogroup 2 had higher troponin T levels (0.01 ng/mL vs 0.02 ng/mL; *P*<.001), indicating that there were more patients in phenogroup 2 who underwent myocardial damage. Patients in phenogroup 2 had higher values of gamma-glutamyl transferase (31.70 IU/L vs 40.30 IU/L; *P*<.001), total bilirubin (12.79 μmol/L vs 15.85 μmol/L; *P*<.001), and aspartate aminotransferase (19.60 IU/L vs 24.29 IU/L; *P*<.001), indicating that patients in phenogroup 2 might have worse liver function compared with phenogroup 1. Moreover, although we had excluded patients with renal dysfunction in advance, patients in phenogroup 2 had worse eGFR values (92.06 mL/min/1.73m^2 vs 81.85 mL/min/1.73 m^2; *P*<.001) and urea (5.46 mmol/L vs mmol/L; *P*<.001). These findings demonstrated that patients in phenogroup 2 had relatively worse kidney function. Furthermore, patients in phenogroup 2 used less angiotensin-converting enzyme inhibitor/angiotensin receptor blocker (phenogroup 1: 1531/2823, 54.23%; phenogroup 2: 1016/2252, 45.11%; *P*<.001), calcium channel blocker (phenogroup 1: 789/2823, 27.95%; phenogroup 2: 321/2252, 14.25%; *P*<.001), and antiplatelets (phenogroup 1: 1914/2823, 67.80%; phenogroup 2: 1384/2823, 61.45%; *P*<.001). It was worth nothing that patients in phenogroup 2 had higher lipid levels (low-density lipoprotein cholesterol and triglyceride) and BMI (25.88 kg/m^2 vs 23.05 kg/m^2; *P*<.001). Compared to phenogroup 1, phenogroup 2 also received less angiography (phenogroup 1: 1311/2823, 46.44%; phenogroup 2: 1311/2252, 30.95%; *P*<.001) and percutaneous coronary intervention (phenogroup 1: 640/2823, 21.96%; phenogroup 2: 349/2252, 15.50%; *P*<.001). Comprehensive baseline characteristics including all 58 candidate variables are listed in Table S5, Multimedia Appendix 1.

Table 2. Baseline characteristics of the PLA General Hospital data set and the generated phenogroups.

Feature	Population (N=5075)	Phenogroup 1 (n=2823)	Phenogroup 2 (n=2252)	*P* value
Feature of interest, n (%)				
AKI[a]	365 (7.19)	106 (3.75)	259 (11.50)	<.001
In-hospital mortality,	139 (2.74)	21 (0.74)	118 (5.24)	<.001
Demographic				
Age (years), median (IQR)	61 (51-70)	58 (48-67)	65 (55-75)	<.001
BMI (kg/m^2), median (IQR)	24.60 (22.46-27.08)	25.88 (23.87-28.08)	23.05 (20.95-25.01)	<.001
DBP[b] (mmHg), median (IQR)	74 (67-81)	77 (70-85)	70 (64-78)	<.001
SBP[c] (mmHg), median (IQR)	125 (113-138)	130 (119-143)	118 (106-130)	<.001
Male, n (%)	3431 (67.61)	1943 (68.83)	1488 (66.07)	<.001
Disease, n (%)				
HF[d]				
Acute HF	1723 (33.95)	738 (26.14)	985 (43.73%)	<.001
Chronic HF	3352 (66.05)	2075 (73.86)	1267 (56.26%)	<.001
AF[e]	1121 (22.09)	526 (18.63)	595 (26.42)	<.001
Cardiomyopathy	941 (18.54)	497 (17.61)	444 (19.71)	<.001
CHD[f]	2928 (57.69)	1660 (58.80)	1268 (56.30)	.07
Diabetes	2002 (39.44)	1041 (36.88)	961 (42.67)	<.001
Stroke	485 (9.56)	282 (9.99)	233 (10.35)	.09
VHD[g]	616 (12.13)	336 (11.90)	280 (12.43)	.57
Medication, n (%)				
ACEI/ARB[h]	2547 (50.18)	1531 (54.23)	1016 (45.11)	<.001
Anticoagulant	1927 (37.97)	989 (35.03)	938 (41.65)	<.001
Antiplatelet	3298 (64.99)	1914 (67.80)	1384 (61.45)	<.001
Beta blocker	3428 (67.54)	1981 (70.17)	1447 (64.25)	<.001
CCB[i]	1110 (21.87)	789 (27.95)	321 (14.25)	<.001
Diuretic	3407 (67.13)	1608 (56.96)	1799 (79.88)	<.001
Positive inotropic drugs	1867 (36.79)	778 (27.56)	1089 (48.36)	<.001
Vasodilator	3103 (61.14)	1698 (60.15)	1405 (62.39)	.10
Echocardiogram				
LVEF[j], median (IQR)	46 (35-56)	50 (39-58)	41 (31-54)	<.001
<40%, n (%)	1716 (33.81)	719 (25.47)	997 (44.27)	<.001
40%-50%, n (%)	1174 (23.13)	690 (24.44)	484 (21.49)	.05
≥50%, n (%)	2185 (42.86)	1414 (50.09)	771 (34.24)	<.001
Laboratory result, median (IQR)				
ALT[k], (IU/L)	21.00 (14.39-33.79)	20.80 (14.70-31.99)	21.54 (13.80-36.49)	<.001
AST[l], (IU/L)	21.29 (16.29-30.50)	19.60 (15.50-26.00)	24.29 (18.09-38.80)	<.001
Calcium (mmol/L)	2.24 (2.16-2.33)	2.28 (2.21-2.36)	2.19 (2.10-2.27)	<.001
eGFR[m] (mL/min/1.73 m^2)	87.62 (75.65-98.80)	92.06 (80.84-101.91)	81.85 (70.90-92.91)	<.001
GGT[n] (IU/L)	34.80 (21.90-63.79)	31.70 (21.30-54.89)	40.30 (23.09-75.00)	<.001

Feature	Population (N=5075)	Phenogroup 1 (n=2823)	Phenogroup 2 (n=2252)	*P* value
HDL-C[o] (mmol/L)	1.02 (0.85-1.22)	1.04 (0.88-1.22)	1.01 (0.82-1.22)	<.001
Hemoglobin, g/L	137 (124-150)	143 (132-154)	129 (116-142)	<.001
LDL-C[p] (mmol/L)	2.25 (1.79-2.81)	2.46 (1.96-3.05)	2.04 (1.62-2.48)	<.001
NT-pro-BNP[q] (pg/mL)	1216 (422-2950)	572 (225-1319)	2680 (1355-5188)	<.001
Potassium (mmol/L)	3.89 (3.62-4.17)	3.87 (3.62-4.13)	3.91 (3.61-4.20)	.005
Sodium (mmol/L)	140.70 (138.10-142.70)	141.30 (139.40-143.20)	139.40 (136.30-142.00)	<.001
Total bilirubin (μmol/L)	13.69 (9.80-19.90)	12.79 (9.40-17.40)	15.85 (10.39-24.60)	<.001
Total protein (g/L)	67.5 (63.3-71.8)	69.2 (65.8-73.3)	65.1 (60.4-69.0)	<.001
Triglyceride (mmol/L)	1.11 (0.82-1.59)	1.34 (0.98-1.87)	0.92 (0.72-1.21)	<.001
Troponin T (ng/mL)	0.01 (0.01-0.04)	0.01 (0.00-0.02)	0.02 (0.01-0.10)	<.001
Urea (mmol/L)	5.84 (4.73-7.25)	5.46 (4.51-6.60)	6.45 (5.11-8.12)	<.001
Operation, n (%)				
Angiography	2008 (29.57)	1311 (46.44)	697 (30.95)	<.001
PCI[r]	969 (19.09)	620 (21.96)	349 (15.50)	<.001

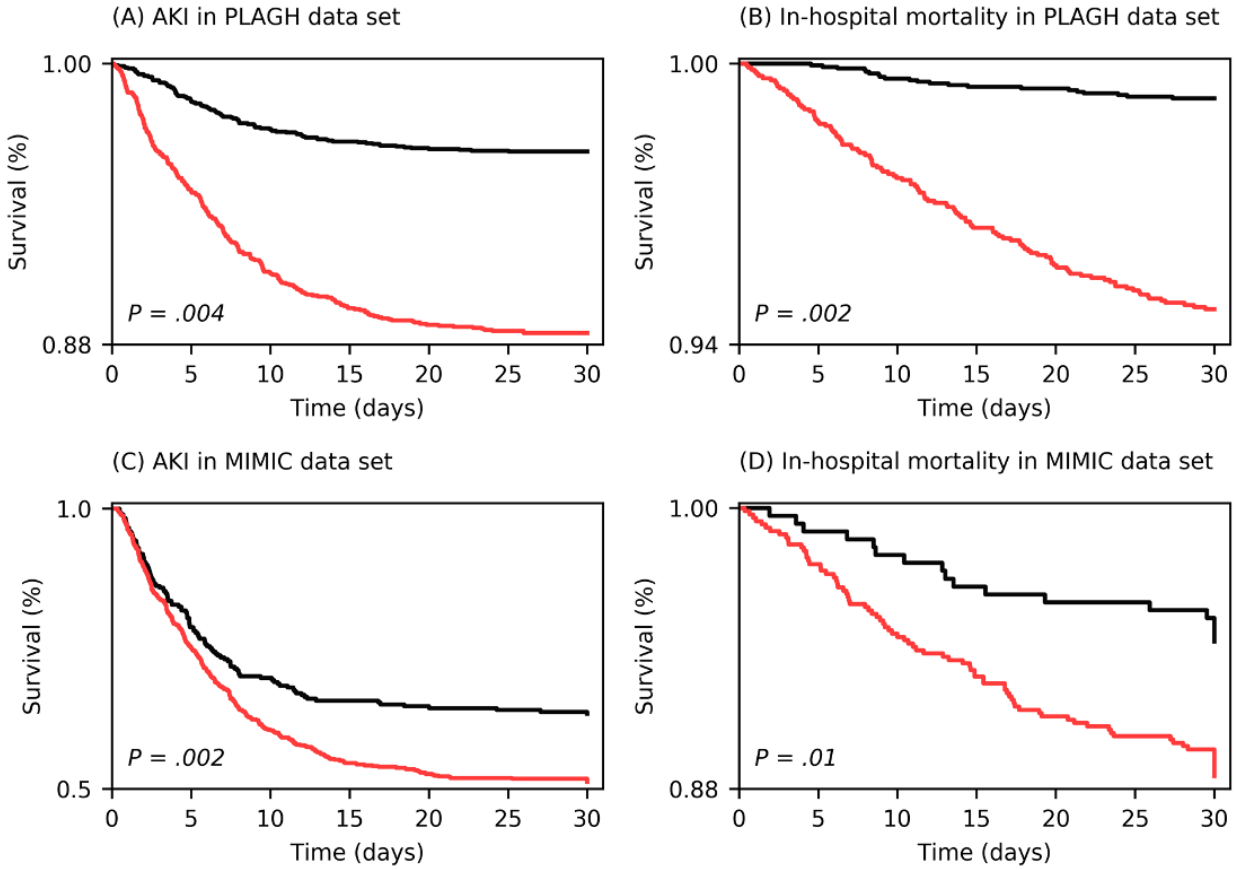
[a]AKI: acuted kidney injury.
[b]DBP: diastolic blood pressure.
[c]SBP: systolic blood pressure.
[d]HF: heart failure.
[e]AF: atrial fibrillation.
[f]CHD: coronary artery disease.
[g]VHD: valvular heart disease.
[h]ACEI/ARB: angiotensin-converting enzyme inhibitor/angiotensin receptor blocker.
[i]CCB: calcium channel blocker.
[j]LVEF: left ventricular ejection fraction.
[k]ALT: alanine aminotransferase.
[l]AST: aspartate transaminase.
[m]eGFR: estimated glomerular filtration rate.
[n]GGT: gamma-glutamyl transferase.
[o]HDL-C: high-density lipoprotein cholesterol.
[p]LDL-C: low-density lipoprotein cholesterol.
[q]NT-pro-BNP: N-terminal probrain natriuretic peptide.
[r]PCI: percutaneous coronary intervention.

Survival Analysis

As the prevalence of AKI and in-hospital mortality had a significant difference between the generated phenogroups, phenogroup 1 was intuitively labeled as “low-risk” and phenogroup 2 as “high-risk.” We further investigated whether the generated phenogroup index could serve as an essential risk indicator for clinical outcomes of interest.

Figure 2 shows the survival difference with respect to AKI and in-hospital mortality between the generated “high-risk” and “low-risk” phenogroups from both the PLAGH data set and the external validation MIMIC-III data set. For AKI, the curves of phenogroup 2 were lower than the curves of phenogroup 1 in both development and external validation data sets (PLAGH: *P*=.004; MIMIC-III: *P*=.002). In addition, we found that most AKI events often happened in the first few days of hospitalization in both the PLAGH and MIMIC-III data sets. This finding was in line with the literature [7,8]. For in-hospital mortality, the curves of phenogroup 2 were consistently lower than the curves of phenogroup 1 (PLAGH: *P*=.002; MIMIC-III: *P*=.01). In consideration of the baseline difference between the PLAGH data set and MIMIC-III data set, the results demonstrated that our model was robust in discriminating between high-risk and low-risk patients and easily transferable to different clinical settings.

Figure 2. Kaplan-Meier curves for AKI and in-hospital mortality in the development (PLAGH) and external validation (MIMIC-III) data sets. AKI: acute kidney injury; MIMIC: Medical Information Mart for Intensive Care; PLAGH: PLA General Hospital.



Outcome Prediction

Table 3 compares the prediction performances of the 5 LR models. Sensitivity, specificity, and concordance statistics are reported for the prediction performance evaluation. As the false-negative prediction (ie, neglecting AKI) may lead to extremely negative consequences, we mainly compared the sensitivity performance among the 5 models. The threshold of sensitivity and specificity was 0.5 in all experiments, and the selected top-10 variables are listed in Table S6, Multimedia Appendix 1. The results showed that the phenogroup index was an essential risk predictor of outcomes. For one, Model 1 used 1 variable (the phenogroup index) as the predictor and achieved promising sensitivity in terms of AKI (0.710) and in-hospital mortality (0.820) among the 5 prediction models with the PLAGH data set. For another, the prediction performance of Model 1 remained quite stable in the external validation (AKI sensitivity 0.760; in-hospital mortality sensitivity 0.826), while there existed significant degradation of performance in the other prediction models.

Table 3. Prediction performance comparison.

Model by task	PLAGH[a] data set (development)			MIMIC-III[b] data set (validation)		
	Sensitivity	Specificity	C-statistic[c]	Sensitivity	Specificity	C-statistics
AKI[d]						
Model 1	0.710	0.577	0.643	0.760	0.342	0.551
Model 2	0.647	0.638	0.696	0.374	0.652	0.532
Model 3	0.679	0.723	0.756	0.478	0.562	0.546
Model 4	0.737	0.753	0.815	0.544	0.560	0.570
Model 5	0.718	0.746	0.816	0.573	0.540	0.575
In-hospital mortality						
Model 1	0.849	0.568	0.708	0.826	0.309	0.568
Model 2	0.791	0.736	0.824	0.530	0.672	0.622
Model 3	0.820	0.763	0.856	0.622	0.599	0.647
Model 4	0.835	0.809	0.899	0.490	0.746	0.646
Model 5	0.856	0.812	0.900	0.620	0.720	0.644

[a]PLAGH: PLA General Hospital.

[b]MIMIC-III: Medical Information Mart for Intensive Care III.

[c]C-statistic: concordance statistic.

[d]AKI: acute kidney injury.

HR Comparison

We used unadjusted Cox proportional hazard regression to determine whether the phenogroup index can act as an essential risk stratification indicator in comparison with the original 39 included variables. The top-ranked 10 variables with the highest HR are listed in Figure 3 (full list is available from Figure S3, Multimedia Appendix 1). The results showed that the HR of the phenogroup index was ranked second in AKI analysis and first in in-hospital mortality analysis, indicating that the phenogroup index can be an effective risk stratification indicator compared with the original variables. Of further note, although troponin T was ranked first for AKI analysis, it was not appropriate for univariate risk indicators since only 16.73% (849/5075) of patients in the PLAGH data set had abnormal records in troponin T. Using troponin T as the indicator only achieved a sensitivity of 0.431, which was significantly lower than the performance of the phenogroup index (0.710). The association between the generated phenogroup index and risk of AKI (in-hospital mortality) was consistent in all examined subgroups (Figure 4).

Figure 3. Hazard ratios of of top-ranked 10 discriminative features for (a) acute kidney injury and (b) in-hospital mortality from the PLA General Hospital data set. AST: aspartate aminotransferase; eGFR: estimated glomerular filtration rate; NT-pro-BNP: N-terminal probrain natriuretic peptide. *Anemia was defined as hemoglobin <135 g/L for men and hemoglobin <120 g/L for women. All units of variables in this figure are same as the units in Table 2.

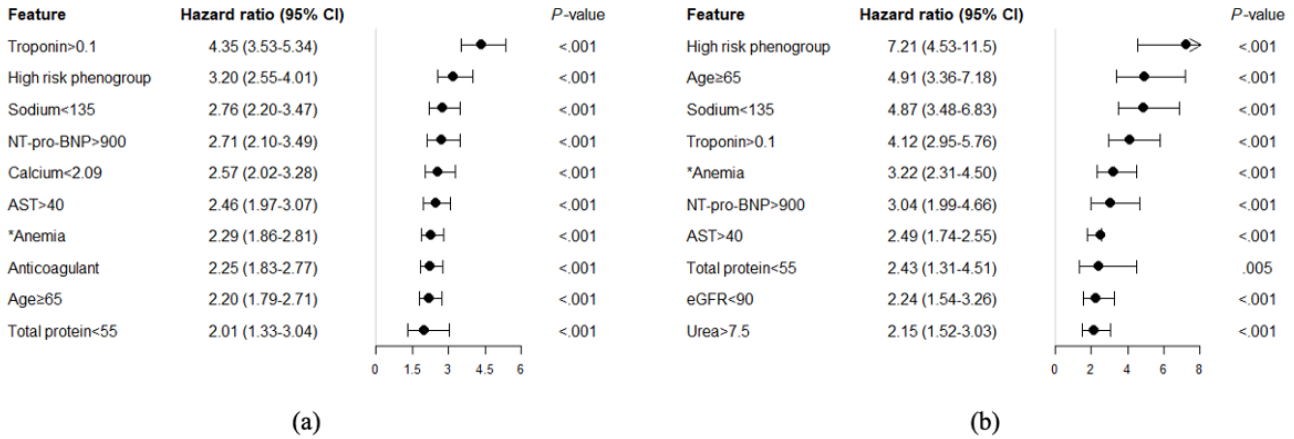
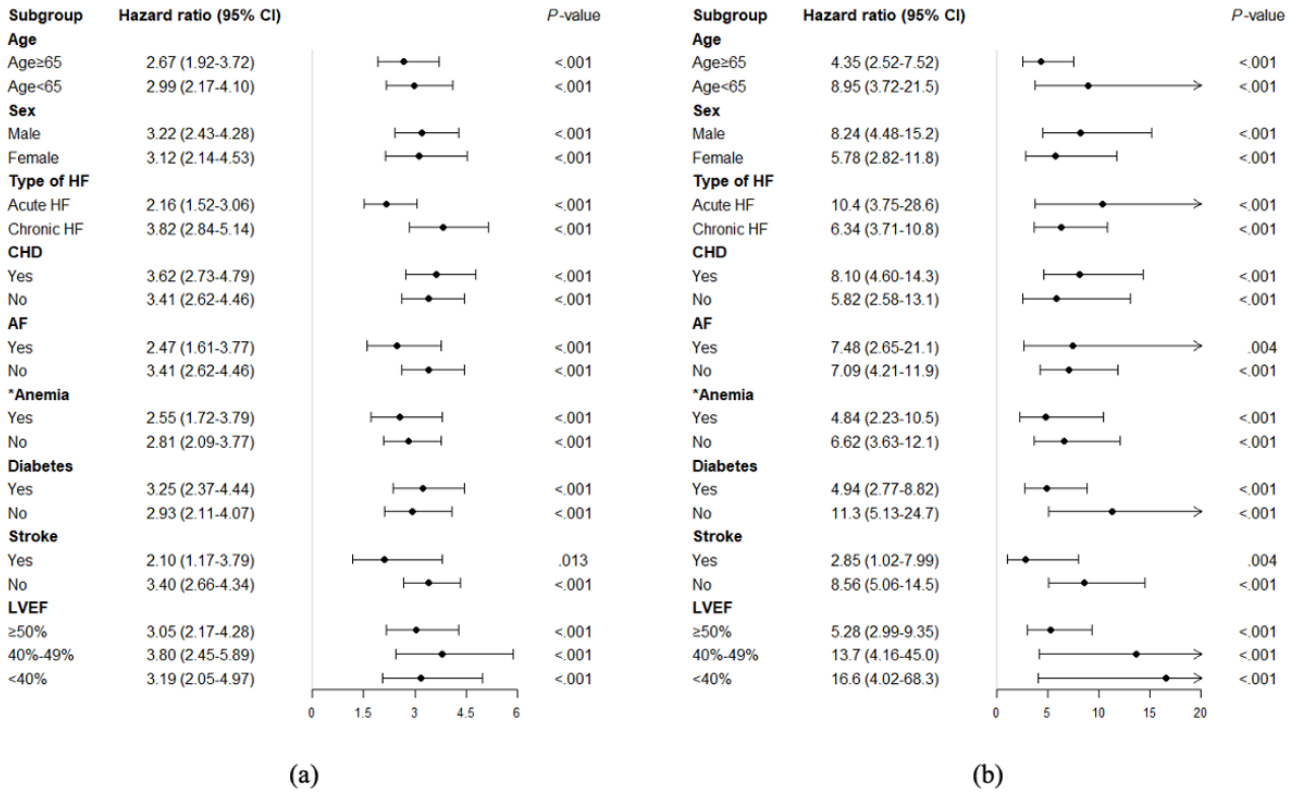


Figure 4. Subgroup analysis of the generated phenogroup index for (a) acute kidney injury and (b) in-hospital mortality. AF: atrial fibrillation; CHD: chronic heart disease; HF: heart failure; LVEF: left ventricular ejection fraction. *Anemia was defined as hemoglobin <135 g/L for men and hemoglobin <120 g/L for women. All units of variables in this figure are same as the units in Table 2.



Discussion

Principal Findings

We explored the potential of using a large volume of EHR data to cluster patients with HF and identify those with normal renal function but susceptible to de novo AKI via an unsupervised machine learning model. The experimental results showed that there was significant difference in AKI and in-hospital mortality occurrence between the 2 phenogroups generated from EHR data. As EHR is a real-world, readily available data source containing rich medical information of thousands of patients, our study demonstrated that it was possible for researchers to answer important clinical and scientific questions effectively by exploiting the huge potential of EHR data via machine learning techniques at a fraction of the resource cost that would have been required using traditional approaches [22,23].

We demonstrated that HF patients with normal renal function can be naturally separated into a “high-risk phenogroup,” of patients susceptible to de novo AKI and a “low-risk phenogroup” who were not. Patients in high-risk phenogroup were typically older, more susceptible to multi-organ dysfunction and anemia, and had significantly higher in-hospital mortality than did those in the low-risk phenogroup. These findings were in line with recent studies [17,24] and warrant further assessment. We found that patients in the high-risk phenogroup had lower levels of lipid and BMI than did those in the low-risk group. These findings are consistent with previous studies reporting that worse cardiac function may cause malnutrition [25] and a decrease of lipid level [26]. Of note, worse cardiac function was also associated with hemodynamic instability, which influences the choice of oral medication strategies [27]. We observed that patients in the high-risk phenogroup received less medication (angiotensin-converting enzyme inhibitor, angiotensin receptor blocker, calcium channel blocker, and beta blockers) than did those in the low-risk phenogroup. On the contrary, we found that patients in low-risk phenogroup were likely to receive percutaneous coronary intervention (PCI) during their stay at the emergency care unit or in hospitalization to revascularize the stable hemodynamic level such that the perfusion of the kidney could be improved and the risk of AKI significantly alleviated. This finding is consistent with previous findings, emphasizing the benefit of timely revascularization [28].

Identification of patients with HF with normal renal function but at high-risk of de novo AKI is a major challenge in HF treatment management. Clinicians have highlighted the need for more effective methods to perform this important clinical task [29]. In this study, we illustrated that machine learning analysis can tackle this challenge by providing deep integration of the comprehensive clinical variables routinely documented in EHR data. As observed in the present study, the phenogroup index generated by an unsupervised machine learning approach, as a latent representation of 39 original variables and their interactions, exhibited a sensitivity of 0.710 and 0.760 on the development data set (PLAGH) and the external validation data set (MIMIC-III). In this sense, the generated phenogroups from raw EHR data are meaningful and can be translated into actionable information for clinical decision-making. On the

contrary, all other LR models met a serious overfitting problem due to the fact that the included variables had different distributions between the development (PLAGH) and external validation (MIMIC-III) data sets (as can be seen in Table S3, Multimedia Appendix 1). Inevitably, this issue caused a significant performance degeneration in the external validation. In consideration of the baseline difference between the PLAGH data set and the MIMIC-III data set, the results suggested that the generated phenogroup index was able to act as an essential de novo AKI risk indicator for patients with HF and normal renal function and be smoothly applied in different clinical settings and in different patient populations. In fact, machine learning algorithms can handle a large volume of variables and a vast number of variable-variable interactions in each patient. This merit effectively individualizes risk assessment and remedies many of the limitations of standard statistical models [22].

Our study has potentially important clinical ramifications. For one, as AKI risk is often underestimated or neglected in patient with HF, especially those with normal renal function [5], our study provided a new perspective for identifying patients with HF and normal renal function but who are at high risk of AKI. For another, in comparison with recent studies that focused on finding new biomarkers for AKI prediction or detection [30], we adopted an improved alternative strategy that used machine learning techniques to explore readily available clinical data to identify patients with HF at high risk of de novo AKI. Such meaningful use of EHR data may provide the best available evidence to assist clinical decision-making. It should be noted that these improvements may be enhanced by mining a large volume of readily available EHR data, which in turn may provide a new avenue for improving any given machine learning algorithm.

Limitations

Several limitations of this study should be acknowledged. First, this is a single-institution study. Although we have evaluated our model on an external validation data set extracted from MIMIC-III, the methods may perform less well in other situations due to the lack of sufficient external validation samples collected from different medical facilities and in different clinical settings. Second, our study was limited by its retrospective design, and all analyses were purely observational. Although we found that there were distinct variables associated with increased risks of de novo AKI and in-hospital mortality, these nonrandomized comparisons should be interpreted cautiously in this context, and the prognostic ability of our model needs to be supported by validation in prospective studies. Third, considering the sensitivity and the specificity for AKI forecasting, our model was relatively sensitive but not very specific. Despite the influence of false-positive classification being limited in this study, further study will be required to enable machine learning–based analysis to capture the salient features distinguishing high- from low-risk cases, such that the prediction performance of our model can be improved.

Conclusions

This study demonstrated that unsupervised machine learning–based EHR analysis is able to separate patients with HF and normal renal function into mutually exclusive phenogroups that correspond to saliently distinct AKI risk levels. Our investigation paves the way for developing an easy-to-use, broadly available model that allows the identification of patients with HF at high-risk of de novo AKI and may help improve outcomes in HF, offering a crucial advantage over traditional techniques for patient phenogrouping and clinical risk stratification.

Acknowledgments

The contribution of investigators and clinical coordinators are duly acknowledged.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Experimental data set introduction and detailed experiment results.
[DOC File , 2618 KB - medinform_v10i10e37484_app1.doc]

References

1. Roy AK, Mc Gorrian C, Treacy C, Kavanaugh E, Brennan A, Mahon NG, et al. A comparison of traditional and novel definitions (RIFLE, AKIN, and KDIGO) of acute kidney injury for the prediction of outcomes in acute decompensated heart failure. Cardiorenal Med 2013 Apr;3(1):26-37 [FREE Full text] [doi: 10.1159/000347037] [Medline: 23801998]
2. Holgado JL, Lopez C, Fernandez A, Sauri I, Uso R, Trillo JL, et al. Acute kidney injury in heart failure: a population study. ESC Heart Fail 2020 Apr;7(2):415-422 [FREE Full text] [doi: 10.1002/ehf2.12595] [Medline: 32059081]
3. Thakar CV, Parikh PJ, Liu Y. Acute kidney injury (AKI) and risk of readmissions in patients with heart failure. Am J Cardiol 2012 May 15;109(10):1482-1486. [doi: 10.1016/j.amjcard.2012.01.362] [Medline: 22381163]
4. Yang L, Xing G, Wang L, Wu Y, Li S, Xu G, et al. Acute kidney injury in China: a cross-sectional survey. The Lancet 2015 Oct;386(10002):1465-1471. [doi: 10.1016/s0140-6736(15)00344-x]

5. Tang X, Chen D, Yu S, Yang L, Mei C, ISN AKF 0 by 25 China Consortium. Acute kidney injury burden in different clinical units: Data from nationwide survey in China. PLoS One 2017 Feb 2;12(2):e0171202 [FREE Full text] [doi: 10.1371/journal.pone.0171202] [Medline: 28152018]
6. Rangaswami J, Bhalla V, Blair JE, Chang TI, Costa S, Lentine KL, American Heart Association Council on the Kidney in Cardiovascular DiseaseCouncil on Clinical Cardiology. Cardiorenal syndrome: classification, pathophysiology, diagnosis, and treatment strategies: a scientific statement from the American Heart Association. Circulation 2019 Apr 16;139(16):e840-e878. [doi: 10.1161/CIR.0000000000000664] [Medline: 30852913]
7. Inohara T, Kohsaka S, Miyata H, Ueda I, Maekawa Y, Fukuda K, et al. Performance and validation of the U.S. NCDR acute kidney injury prediction model in Japan. J Am Coll Cardiol 2016 Apr 12;67(14):1715-1722 [FREE Full text] [doi: 10.1016/j.jacc.2016.01.049] [Medline: 27056778]
8. Abusaada K, Yuan C, Sabzwari R, Butt K, Maqsood A. Development of a novel score to predict the risk of acute kidney injury in patient with acute myocardial infarction. J Nephrol 2017 Jun;30(3):419-425. [doi: 10.1007/s40620-016-0326-1] [Medline: 27300206]
9. Cikes M, Sanchez-Martinez S, Claggett B, Duchateau N, Piella G, Butakoff C, et al. Machine learning-based phenogrouping in heart failure to identify responders to cardiac resynchronization therapy. Eur J Heart Fail 2019 Jan;21(1):74-85 [FREE Full text] [doi: 10.1002/ejhf.1333] [Medline: 30328654]
10. Shah SJ, Katz DH, Selvaraj S, Burke MA, Yancy CW, Gheorghiade M, et al. Phenomapping for novel classification of heart failure with preserved ejection fraction. Circulation 2015 Jan 20;131(3):269-279. [doi: 10.1161/circulationaha.114.010637]
11. Segar MW, Patel KV, Ayers C, Basit M, Tang WW, Willett D, et al. Phenomapping of patients with heart failure with preserved ejection fraction using machine learning-based unsupervised cluster analysis. Eur J Heart Fail 2020 Jan;22(1):148-158 [FREE Full text] [doi: 10.1002/ejhf.1621] [Medline: 31637815]
12. Levey AS, Stevens LA, Schmid CH, Zhang Y, Castro AF, Feldman HI, CKD-EPI (Chronic Kidney Disease Epidemiology Collaboration). A new equation to estimate glomerular filtration rate. Ann Intern Med 2009 May 05;150(9):604-612 [FREE Full text] [doi: 10.7326/0003-4819-150-9-200905050-00006] [Medline: 19414839]
13. Buuren SV, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software 2011 Dec 12;45(3):1-67. [doi: 10.18637/jss.v045.i03]
14. James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning. New York, NY: Springer; 2013.
15. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 1987 Nov;20:53-65. [doi: 10.1016/0377-0427(87)90125-7]
16. van der Maaten L, Hinton G. Visualizing data using t-SNE. Journal of Machine Learning Research 2008;9:2579-2605.
17. Khwaja A. KDIGO clinical practice guidelines for acute kidney injury. Nephron Clin Pract 2012;120(4):c179-c184 [FREE Full text] [doi: 10.1159/000339789] [Medline: 22890468]
18. Johnson AE, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016 May 24;3:160035 [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
19. Abraham A, Pedregosa F, Eickenberg M, Gervais P, Mueller A, Kossaifi J, et al. Machine learning for neuroimaging with scikit-learn. Front Neuroinform 2014;8:14 [FREE Full text] [doi: 10.3389/fninf.2014.00014] [Medline: 24600388]
20. Davidson-Pilon C. lifelines: survival analysis in Python. JOSS 2019 Aug;4(40):1317. [doi: 10.21105/joss.01317]
21. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Polat, SciPy 1.0 Contributors. Author Correction: SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 2020 Mar;17(3):352 [FREE Full text] [doi: 10.1038/s41592-020-0772-5] [Medline: 32094914]
22. Diller GP, Kempny A, Babu-Narayan SV, Henrichs M, Brida M, Uebing A, et al. Machine learning algorithms estimating prognosis and guiding therapy in adult congenital heart disease: data from a single tertiary centre including 10 019 patients. Eur Heart J 2019 Apr 01;40(13):1069-1077 [FREE Full text] [doi: 10.1093/eurheartj/ehy915] [Medline: 30689812]
23. Tokodi M, Schwertner W, Kovács A, Tősér Z, Staub L, Sárkány A, et al. Machine learning-based mortality prediction of patients undergoing cardiac resynchronization therapy: the SEMMELWEIS-CRT score. Eur Heart J 2020 May 07;41(18):1747-1756 [FREE Full text] [doi: 10.1093/eurheartj/ehz902] [Medline: 31923316]
24. Zymliński R, Sokolski M, Biegus J, Siwołowski P, Nawrocka-Millward S, Sokolska JM, et al. Multi-organ dysfunction/injury on admission identifies acute heart failure patients at high risk of poor outcome. Eur J Heart Fail 2019 Jun;21(6):744-750 [FREE Full text] [doi: 10.1002/ejhf.1378] [Medline: 30561066]
25. Sze S, Pellicori P, Zhang J, Clark AL. Malnutrition, congestion and mortality in ambulatory patients with heart failure. Heart 2019 Feb;105(4):297-306. [doi: 10.1136/heartjnl-2018-313312] [Medline: 30121635]
26. Pitt B, Loscalzo J, Ycas J, Raichlen JS. Lipid levels after acute coronary syndromes. J Am Coll Cardiol 2008 Apr 15;51(15):1440-1445 [FREE Full text] [doi: 10.1016/j.jacc.2007.11.075] [Medline: 18402897]
27. Ponikowski P, Voors AA, Anker SD, Bueno H, Cleland JG, Coats AJ, et al. 2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure. Kardiol Pol 2016 Oct 13;74(10):1037-1147. [doi: 10.5603/kp.2016.0141]
28. Shacham Y, Steinvil A, Arbel Y. Acute kidney injury among ST elevation myocardial infarction patients treated by primary percutaneous coronary intervention: a multifactorial entity. J Nephrol 2016 Apr;29(2):169-174. [doi: 10.1007/s40620-015-0255-4] [Medline: 26861658]

29. Abusaada K, Yuan C, Sabzwari R, Butt K, Maqsood A. Development of a novel score to predict the risk of acute kidney injury in patient with acute myocardial infarction. J Nephrol 2017 Jun;30(3):419-425. [doi: 10.1007/s40620-016-0326-1] [Medline: 27300206]
30. Lassus JPE, Nieminen MS, Peuhkurinen K, Pulkki K, Siirilä-Waris K, Sund R, FINN-AKVA study group. Markers of renal function and acute kidney injury in acute heart failure: definitions and impact on outcomes of the cardiorenal syndrome. Eur Heart J 2010 Nov 27;31(22):2791-2798. [doi: 10.1093/eurheartj/ehq293] [Medline: 20801926]

Abbreviations

AF: atrial fibrillation
AKI: acute kidney injury
CHD: coronary heart disease
CKD: chronic kidney disease
CKD-EPI: Chronic Kidney Disease Epidemiology Collaboration
eGFR: estimated glomerular filtration rate
EHR: electronic health record
HF: heart failure
HR: hazard ratio
KDIGO: Kidney Disease: Improving Global Outcomes
LR: logistic regression
MIMIC: Medical Information Mart for Intensive Care
PLAGH: Chinese PLA General Hospital
SCr: serum creatinine

Edited by T Hao, B Tang, Z Li; submitted 23.02.22; peer-reviewed by H Monday, G Nneji, A Naser, K Uludag; comments to author 15.05.22; revised version received 31.05.22; accepted 05.06.22; published 14.10.22.

Please cite as:
Hong C, Sun Z, Hao Y, Dong Z, Gu Z, Huang Z
Identifying Patients With Heart Failure Who Are Susceptible to De Novo Acute Kidney Injury: Machine Learning Approach
JMIR Med Inform 2022;10(10):e37484
URL: https://medinform.jmir.org/2022/10/e37484
doi:10.2196/37484
PMID:36240002

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org